Ian F. C. Smith
Bernd Domer (Eds.)

LNCS 10863

Advanced Computing Strategies for Engineering

25th EG-ICE International Workshop 2018
Lausanne, Switzerland, June 10–13, 2018
Proceedings, Part I

1 Part I

Springer

Lecture Notes in Computer Science 10863

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, Lancaster, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Zurich, Switzerland

John C. Mitchell
Stanford University, Stanford, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

C. Pandu Rangan
Indian Institute of Technology Madras, Chennai, India

Bernhard Steffen
TU Dortmund University, Dortmund, Germany

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Gerhard Weikum
Max Planck Institute for Informatics, Saarbrücken, Germany

More information about this series at http://www.springer.com/series/7409

Ian F. C. Smith · Bernd Domer (Eds.)

Advanced Computing Strategies for Engineering

25th EG-ICE International Workshop 2018
Lausanne, Switzerland, June 10–13, 2018
Proceedings, Part I

Springer

Editors

Ian F. C. Smith
Applied Computing and Mechanics
Laboratory (IMAC)
School of Architecture, Civil and
Environmental Engineering (ENAC)
Swiss Federal Institute of Technology,
Lausanne (EPFL)
Lausanne
Switzerland

Bernd Domer
Institute for Landscape, Architecture,
Construction and Territory (inPact)
Construction and Environment
Department (CED)
University of Applied Sciences,
Geneva (HEPIA)
Geneva
Switzerland

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-91634-7 ISBN 978-3-319-91635-4 (eBook)
https://doi.org/10.1007/978-3-319-91635-4

Library of Congress Control Number: 2018943445

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer International Publishing AG, part of Springer Nature 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The architecture–engineering–construction (AEC) industry worldwide spending is over ten trillion dollars annually[1]. The industry is the largest global consumer of raw materials, and constructed assets account for 25–49% of the world's total carbon emissions[2]. Also, the World Bank has estimated that each year, demand for civil infrastructure exceeds supply (new plus existing infrastructure) creating an annual shortfall of one trillion dollars[3]. This cannot continue. Engineers must find new ways to design, build, manage, renovate, and recycle buildings and civil infrastructure.

Advanced computing strategies for engineering will be the enablers for much of this transformation. Until recently, new computing strategies have not been able to penetrate into the AEC industry. Owners and other stakeholders have observed little return on investment along with excessive risk associated with a fragmented industry where computing competence is far from homogeneous. This is changing quickly as efficient information modeling, the foundation of many computing strategies in this field, becomes more accessible. Also, important advances in fields such as construction management, life-cycle design, monitoring, diagnostics, asset management, and structural control are being made thanks to fundamental computing advances in fields such as machine learning, model-based reasoning, and human–computer interaction. In parallel, studies of full-scale AEC cases are uncovering additional scientific challenges for computer scientists.

The European Group for Intelligent Computing in Engineering (EG-ICE) was established in Lausanne in 1993 to promote research that lies on the interface between computing and engineering challenges. The primary goals of the group are to promote engineering informatics research across Europe by improving communication and trust between researchers, fostering collaborative research, and enhancing awareness of recent research. The EG-ICE group maintains contact with similar groups outside Europe and encourages contact with experts wherever they reside.

This volume contains papers that were presented at the 25th Workshop of the European Group for Intelligent Computing in Engineering (EG-ICE), which was held in Lausanne, Switzerland, June 10–13, 2018. Of the 108 abstracts that were submitted, 57 papers made it through the multi-step review process of evaluating abstracts, commenting on full papers, and assessing subsequent revisions so that they could be presented at the workshop.

[1] https://www.statista.com/statistics/788128/construction-spending-worldwide/.

[2] Shaping the Future of Construction, World Economic Forum, Geneva, 2016.

[3] https://futureofconstruction.org/blog/infographic-six-megatrends-impacting-the-ec-industry/.

We are grateful to the many reviewers who worked hard to provide constructive comments to authors. The scientific results presented here are a sample of the diversity and creativity of those who are planting the seeds of the exciting transformation that is coming over the next decade. It is not too soon.

April 2018

Ian F. C. Smith
Bernd Domer

Organization

Organizing Committee

Ian Smith (Workshop Co-chair)	Swiss Federal Institute of Technology, EPFL, Switzerland
Bernd Domer (Workshop Co-chair)	University of Applied Sciences, HEPIA, Switzerland
Raphaël Wegmann (Workshop Secretary)	Swiss Federal Institute of Technology, EPFL, Switzerland
Pierino Lestuzzi (Member)	Swiss Federal Institute of Technology, EPFL, Switzerland
Sai Pai (Member)	Swiss Federal Institute of Technology, EPFL, Switzerland
Yves Reuland (Member)	Swiss Federal Institute of Technology, EPFL, Switzerland
Gennaro Senatore (Member)	Swiss Federal Institute of Technology, EPFL, Switzerland
Ann Sychertz (Member)	Swiss Federal Institute of Technology, EPFL, Switzerland

EG-ICE Committee

Pieter de Wilde (Chair)	University of Plymouth, UK
Timo Hartmann (Vice-chair)	Technical University of Berlin, Germany
Haijiang Li (Secretary)	Cardiff University, UK
Philipp Geyer (Treasurer and International Representative)	Catholic University of Leuven, Belgium
Georg Suter (Chair, Best Paper Award Committee)	Technical University of Vienna, Austria
Jakob Beetz (Committee Member)	RWTH Aachen University, Germany
Christian Koch (Committee Member)	Bauhaus University, Weimar, Germany
André Borrmann (Past Chair)	Technical University of Munich, Germany
Ian Smith (EG-ICE Fellow)	Swiss Federal Institute of Technology, EPFL, Switzerland

Scientific Committee

Jamal Abdalla	American University of Sharjah, UAE
Burcu Akinci	Carnegie Mellon University, USA

Robert Amor	University of Auckland, New Zealand
Chimay Anumba	University of Florida, USA
Burcin Becerik-Gerber	University of Southern California, USA
Jakob Beetz	RWTH Aachen University, Germany
Mario Berges	Carnegie Mellon University, USA
André Borrmann	Technical University of Munich, Germany
Frédéric Bosché	Heriot-Watt University, UK
Manfred Breit	University of Applied Sciences, FHNW, Switzerland
Ioannis Brilakis	University of Cambridge, UK
Hubo Cai	Purdue University, USA
Jack Cheng	Hong Kong University of Science and Technology, SAR China
Symeon Christodoulou	University of Cyprus, Cyprus
Lorenzo Diana	Swiss Federal Institute of Technology, EPFL, Switzerland
Semiha Ergan	New York University, USA
Esin Ergen	Istanbul Technical University, Turkey
Boi Faltings	Swiss Federal Institute of Technology, EPFL, Switzerland
Martin Fischer	Stanford University, USA
Ian Flood	University of Florida, USA
Adel Francis	École de Technologie Supérieure, ÉTS, Canada
Renate Fruchter	Stanford University, USA
James Garrett	Carnegie Mellon University, USA
David Gerber	University of Southern California, USA
Philipp Geyer	Catholic University of Leuven, Belgium
Mani Golparvar-Fard	University of Illinois at Urbana-Champaign, USA
Ewa Grabska	Jagiellonian University, Poland
Carl Haas	University of Waterloo, Canada
Amin Hammad	Concordia University, Canada
Timo Hartmann	Technical University of Berlin, Germany
Markku Heinisuo	Tampere University of Technology, Finland
Shang-Hsien (Patrick) Hsieh	National Taiwan University, Taiwan
Raja Raymond Issa	University of Florida, USA
Farrokh Jazizadeh	Virginia Polytechnic Institute and State University, USA
Vineet Kamat	University of Michigan, USA
Peter Katranuschkov	Technical University of Dresden, Germany
Arto Kiviniemi	University of Liverpool, UK
Christian Koch	Bauhaus University, Weimar, Germany
Bimal Kumar	UK
Markus König	Ruhr University of Bochum, Germany
Debra Laefer	New York University, USA
Kincho Law	Stanford University, USA
SangHyun Lee	University of Michigan, USA
Fernanda Leite	University of Texas at Austin, USA
Haijiang Li	Cardiff University, UK
Ken-Yu Lin	University of Washington, USA

Jerome Lynch	University of Michigan, USA
John Messner	Pennsylvania State University, USA
Edmond Miresco	École de Technologie Supérieure, ÉTS, Canada
Ivan Mutis	Illinois Institute of Technology, USA
Hae Young Noh	Carnegie Mellon University, USA
William O'Brien	University of Texas at Austin, USA
Esther Obonyo	Pennsylvania State University, USA
Feniosky Peña-Mora	Columbia University, USA
Yaqub Rafiq	University of Plymouth, UK
Yves Reuland	Swiss Federal Institute of Technology, EPFL, Switzerland
Uwe Rüppel	Technical University of Darmstadt, Germany
Rafael Sacks	Technion - Israel Institute of Technology, Israel
Eduardo Santos	University of Sao Paulo, Brazil
Sergio Scheer	Federal University of Parana, Brazil
Raimar Scherer	Technical University of Dresden, Germany
Gennaro Senatore	Swiss Federal Institute of Technology, EPFL, Switzerland
Kristina Shea	Swiss Federal Institute of Technology, ETH Zurich, Switzerland
Kay Smarsly	Bauhaus University, Weimar, Germany
Lucio Soibelman	University of Southern California, USA
Sheryl Staub-French	University of British Columbia, Canada
Georg Suter	Technical University of Vienna, Austria
Pingbo Tang	Arizona State University, USA
Jochen Teizer	Ruhr University of Bochum, Germany
Walid Tizani	University of Nottingham, UK
Žiga Turk	University of Ljubljana, Slovenia
Xiangyu Wang	Curtin University, Australia
Nobuyoshi Yabuki	University of Osaka, Japan
Yimin Zhu	Louisiana State University, USA
Pieter de Wilde	University of Plymouth, UK

Contents – Part I

Advanced Computing in Engineering

Computer Supported Construction Management

Life-Cycle Design Support

Contents – Part II

Monitoring and Control Algorithms in Engineering

BIM and Engineering Ontologies

Advanced Computing in Engineering

Automatic Object Detection from Digital Images by Deep Learning with Transfer Learning

Nobuyoshi Yabuki, Naoto Nishimura, and Tomohiro Fukuda

Osaka University, Suita, Osaka 565-0871, Japan
yabuki@see.eng.osaka-u.ac.jp

Abstract. At construction sites and disaster areas, an enormous number of digital photographs are taken by engineers. Tasks such as collecting, sorting, annotating, storing, deleting, distributing these digital images, as done manually, are cumbersome, error-prone, and time-consuming. Thus, it is desirable to automate the object detection process of pictures so that engineers do not have to waste their valuable time and can improve the efficiency and accuracy. Although conventional machine learning could be a solution, it takes much time for researchers to determine features and contents of digital images, and the accuracy tends to be unsatisfactory. On the other hand, deep learning can automatically determine features and contents of various objects from digital images. Therefore, this research aims to automatically detect each object as an object and its position from digital images by using deep learning. Since deep learning usually requires a very large amount of dataset, this research has adopted deep learning with transfer learning, which enables object detection even if the dataset is not very large. Experiments were executed to detect construction machines, workers, and signboards in photographs, comparing among the conventional machine learning by feature values, deep learning with and without transfer learning. The result showed that the best performance was achieved by the deep learning with transfer learning.

Keywords: Deep learning · Image detection · Transfer learning

1 Introduction

An enormous number of pictures are taken at construction sites and disaster areas by engineers and are used for inspection, management, disaster recovery, and scientific interests. Those pictures include objects such as construction machinery, signage, signboards, construction workers, etc. In order to use those pictures for inspection and management of construction, those objects must be manually detected by humans. In order to improve the efficiency, the detection process should be automatically executed. Recently, photographs can be classified by object detection functions using machine learning, e.g., People of iOS [1] and face grouping in Google Photo [2]. However, those systems are mostly for ordinary things such as faces, automobiles, bicycles, televisions,

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 3–15, 2018.
https://doi.org/10.1007/978-3-319-91635-4_1

etc., but not for specific civil engineering entities such as construction machinery, signboards at construction sites, stakes, scaffoldings, etc.

In object detection in images, machine learning using features of the image has often been employed. Features used for object detection include local features, such as Haar-like features [3], Histograms of Oriented Gradients (HOG) features [4], Edge Oriented Gradients features [5], Edgelet features [6], etc. However, appropriate features must be determined by the user and it is difficult to achieve the satisfactory level of object detection. In addition, since using single feature makes much false detection, joint features or boosting [7] which combines plural object detection mechanisms are necessary, which requires much work and effort during the machine learning phase.

On the other hand, in deep learning or Convolutional Neural Networks (CNNs), features of objects can automatically be found and computed. Thus, the user does not have to determine the features and the accuracy of object detection has been improved significantly. Research on ordinary object detection include VGG-16 [8], Regions with CNN features (R-CNN) [9], Fast Region-based Convolutional Network (Fast R-CNN) [10], Faster R-CNN [11], Single Shot Multibox Detector (SSD) [12], You Only Look Once (YOLO) [13], etc. However, huge amount of data is required for deep learning. For ordinary object detection, large datasets for learning such as Pascal VOC [14], Caltech101 [15], COCO dataset [16, 17], UCI Machine Learning Repository [18] are provided for learning. For example, Pascal VOC has 20 classes such as person, animal, vehicle, indoor with 9,963 images containing 24,640 annotated objects. However, datasets for non-ordinary objects such as construction machinery, civil engineering objects are not prepared yet and must be developed to achieve the high accuracy in object detection.

Machine learning methods include supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning. Recently, transfer learning receives attention from Artificial Intelligence (AI) researchers. In transfer learning [19, 20], knowledge is stored for solving one problem and it is applied to a different but related problem. Transfer learning is thought to be particularly useful when large dataset cannot be collected easily. Fine tuning is a subset of transfer learning, which classifies objects that are already classified with similar but different labels. In this research, transfer learning but not fine tuning is employed.

The objective of this research is to develop a method to automatically detect construction domain specific objects such as construction machines, construction workers, construction signboards from digital images using deep learning with transfer learning and to classify the pictures according to the image contents. The reason for adopting transfer learning is that existing datasets lack of civil engineering specific objects and that it is hard to collect huge amount of such data to be needed for deep learning in a short time with a limited fund. To verify the proposed method, performance is compared among conventional machine learning using HOG features, deep learning with and without transfer learning. The reason that the single HOG features and Support Vector Machine (SVM) was employed for object detection is that it has an advantage for local shape change over other features described before.

2 Proposed Object Detection Method

2.1 Overview of the Proposed Method

The proposed method is to detect objects of the construction-specific domain and their positions from digital images. SSD is employed to detect positions as well as the types of the objects in images. Next, correct answer labels to re-learn SSD is created. LabelImg [21] is used to create learning data labels. Finally, the created dataset is re-learned to the SSD. By using the new weight acquired by transfer learning, the object position on the digital image are determined.

In this research, Keras was used as a machine learning library. It is one of the machine learning libraries written in Python. TensorFlow can be used as the back end with Keras. TensorFlow, which is developed and provided by Google Inc., was used for recognizing the object position as the backend. Figure 1 shows the overview of the flow of object position detection method.

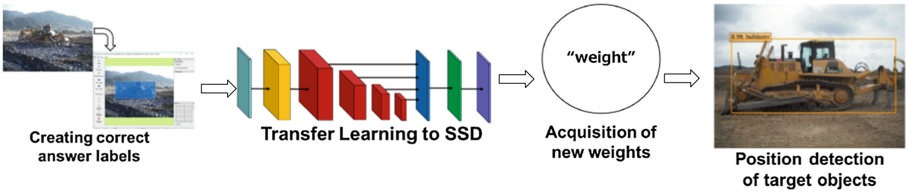

Creating correct answer labels
Transfer Learning to SSD
“weight”
Acquisition of new weights
Position detection of target objects

Fig. 1. Overview of the flow of object position detection method.

SSD is an object detection algorithm using a simple network built by Liu et al. SSD is faster than conventional object detection algorithms. A simplified network model of SSD is shown in Fig. 2. VGG-16, which is a learning model of image detection, is used for the base network of SSD. After this base network, by using hierarchical feature maps, various scale objects can be processed, and the accuracy of detection rate can be high. The reason is that it identifies for each aspect ratio of the object. For these reasons, the

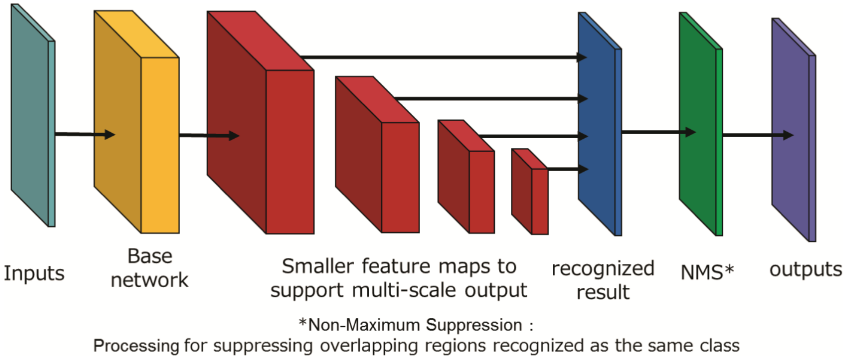
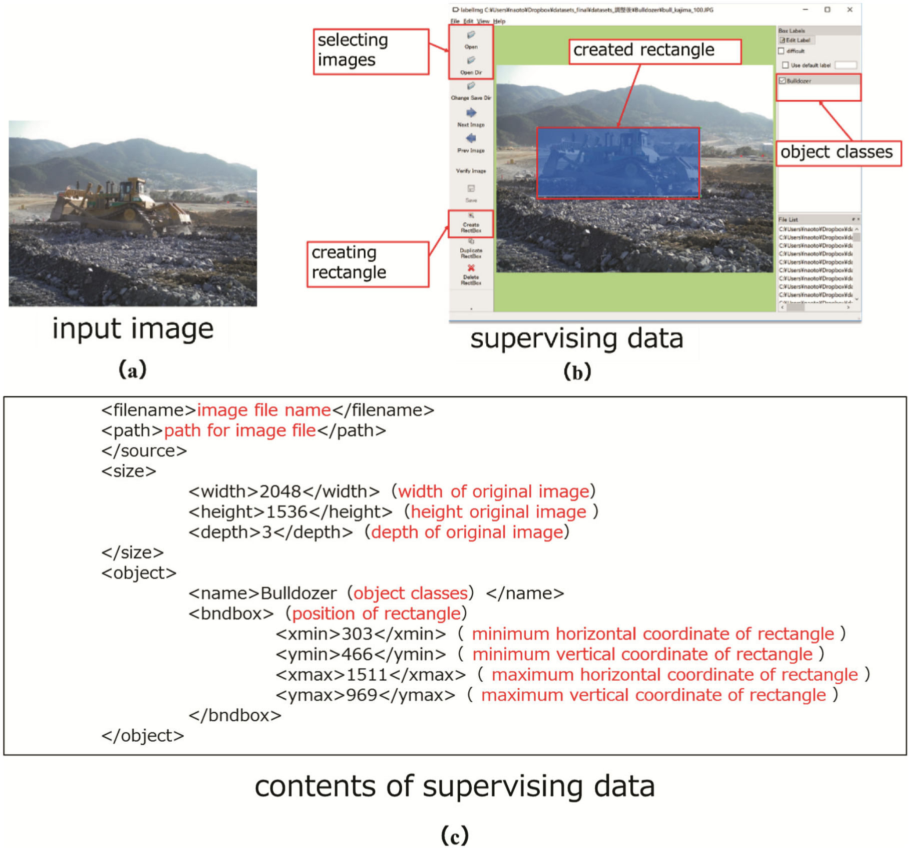

Inputs
Base network
Smaller feature maps to support multi-scale output
recognized result
NMS*
outputs
*Non-Maximum Suppression :
Processing for suppressing overlapping regions recognized as the same class

Fig. 2. Simplified network image model of SSD (drawn by the authors referencing to Fig. 2 of Liu et al.).

SSD can detect a target object even from a relatively low-resolution image. Even if digital images taken at a construction site are high resolution, a low-resolution object exists in the distance of images.

2.2 Detection of Object Type and Position

Creation of Correct Answer Labels for Object Position Detection. In order to detect an object from a digital image with machine learning, it is necessary to know what exists and where it is. Hence, a bounding box and its coordinates are required to create correct answer labels. In addition, correct answer labels must exist with the original digital images. They are created by using LabelImg [21]. This tool was built by Tzutalin to create a correct label called annotation data. The annotation data includes information such as the name of the original image and coordinates of a bounding box (Horizontal and vertical coordinates on the image). First, image data to be learned is opened (Fig. 3(a)). Then, the coordinates of upper left and lower right corners of the object are measured. It is possible to create bounding boxes of a plurality of objects for one image



input image

(a)

supervising data

(b)

```
        <filename>image file name</filename>
        <path>path for image file</path>
        </source>
        <size>
                <width>2048</width> (width of original image)
                <height>1536</height> (height original image )
                <depth>3</depth> (depth of original image)
        </size>
        <object>
                <name>Bulldozer (object classes) </name>
                <bndbox> (position of rectangle)
                        <xmin>303</xmin> ( minimum horizontal coordinate of rectangle )
                        <ymin>466</ymin> ( minimum vertical coordinate of rectangle )
                        <xmax>1511</xmax> ( maximum horizontal coordinate of rectangle )
                        <ymax>969</ymax> ( maximum vertical coordinate of rectangle )
                </bndbox>
        </object>
```

contents of supervising data

(c)

Fig. 3. Sample of supervising data.

(Fig. 3(b)). The created annotation data is saved as an XML format file as shown in Fig. 3(c).

Transfer Learning and Object Type/Position Detection. The created correct labels and the original images are learned and stored in SSD. When learning is completed, a new set of weights is acquired. An appropriate set of weights is selected from the newly acquired weights and used for object detection. The selection of the appropriate set of weights is done from the transition of the loss coefficient acquired by transfer learning (Fig. 4). In Fig. 4, since the difference of loss between training data and test data is minimum at Epoch 41, the weight at this point is selected.

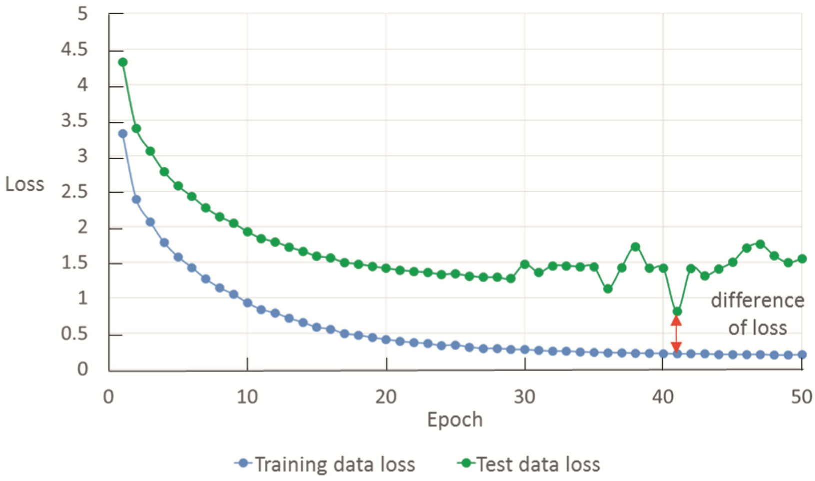


Fig. 4. A sample of transition of training data loss and test data loss.

Classification of Digital Image Files. File folders of which names are detected objects such as bulldozers, backhoes, dump trucks, wheel loaders, workers, signboards, etc. are prepared in a computer. All digital images files that at least one object is detected are put into the designated file folder. If two or more objects are detected, the file is copied and put into the multiple designated folders.

3 Experiments

Two kinds of experiments were executed. Experiment I is to compare the detection result between the proposed method, i.e., deep learning with transfer learning and deep learning without transfer learning. Experiment II is to compare the results among (a) deep learning with transfer learning, (b) deep learning without transfer learning, and (c) conventional machine learning. Construction machinery (backhoes, bulldozers, dump tracks, and wheel loaders), workers, and signboards were selected as objects to be detected, and the accuracy of object position detection was tested. Detection accuracy is evaluated based on the criteria shown in Table 1. System environment for learning and testing is shown in Table 2. The number of images for Experiment I, II, and for testing is shown in Table 3.

Table 1. Evaluation criteria for detection accuracy.

Types of detection	Position detection
Positive detection	Both position and object type are correctly detected
Negative detection	Either (1) position is correctly detected but the object type is incorrect or (2) the object type is correctly detected but the position is incorrect
Not detected	Nothing is detected

Table 2. System environment for learning and testing.

Item	Type
CPU	Intel Core i7-7700 BOX CPU @3.60 GHz
Main memory	64 GB
GPU	GeForce GTX 1080 Ti. Memory: 11 Gbps, 11 GB
OS	Ubuntu 14.04, 64 bit (Linux)
Language	Python 2.0

Table 3. The number of images for Experiment I, II, and testing.

Object	Experiment I	Experiment II	Testing
Backhoe	115	84	202
Bulldozer	113	111	107
Dump truck	100	97	115
Wheel loader	117	117	123
Worker	302	249	64
Signboard	300	217	78

3.1 Experiment I: Verification for Deep Learning with Transfer Learning

Experiment I was executed for comparing the detection accuracy result between deep learning with transfer learning and without it. The model used for deep learning with transfer learning is named I-A while that for without it is named I-B in this research. 90% of the learning dataset was used for training and 10% for testing.

Figure 5 shows samples for positive detection and negative detection cases for each object. For example, in the negative detection of backhoe, the machine was detected as a dump truck. Figure 6 shows the result of the object detection experiment using the testing data. Detection accuracy for I-A (deep learning with transfer learning) was 86.6% for backhoes, 61.7% for bulldozers, 80.9% for dump trucks, 86.2% for wheel loaders, 100.0% for signboards, and 79.7% for workers. On the other hand, accuracy for I-B (deep learning without transfer learning) was, 0.0%, 20.6%, 0.0%, 17.1%, 60.9%, and 0.0%, respectively. Obviously, I-A shows much accuracy rate than I-B.

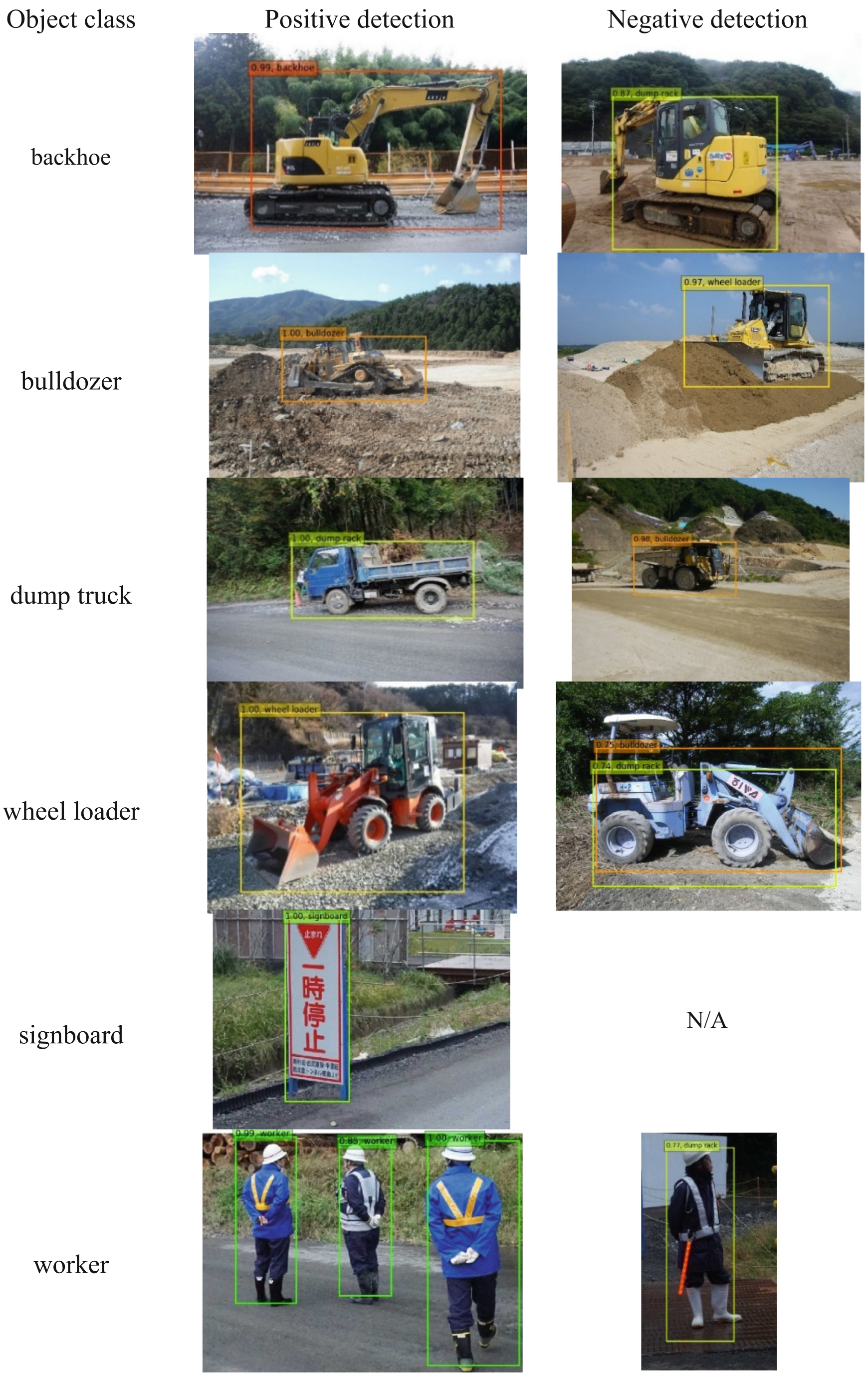
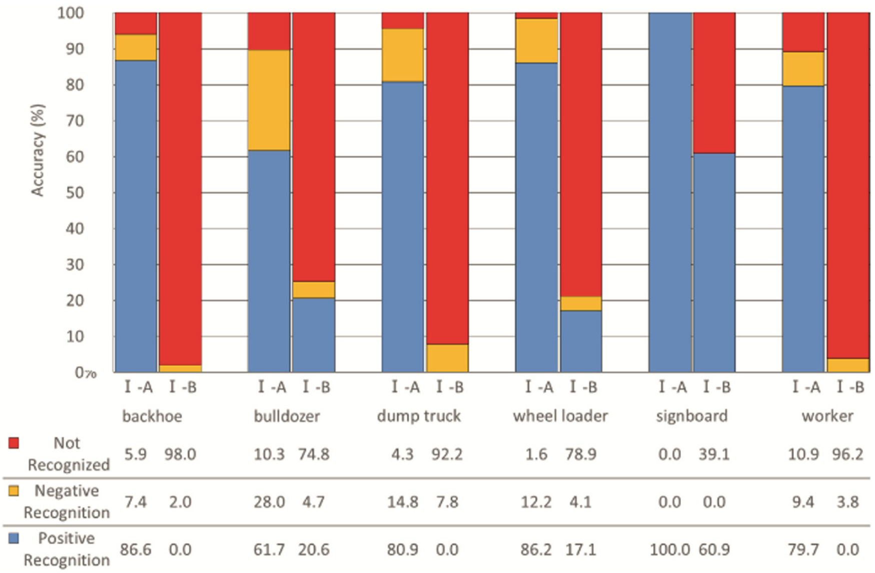


Fig. 5. Samples for positive and negative detection cases for each object class.



	backhoe I -A	backhoe I -B	bulldozer I -A	bulldozer I -B	dump truck I -A	dump truck I -B	wheel loader I -A	wheel loader I -B	signboard I -A	signboard I -B	worker I -A	worker I -B
🟥 Not Recognized	5.9	98.0	10.3	74.8	4.3	92.2	1.6	78.9	0.0	39.1	10.9	96.2
🟨 Negative Recognition	7.4	2.0	28.0	4.7	14.8	7.8	12.2	4.1	0.0	0.0	9.4	3.8
🟦 Positive Recognition	86.6	0.0	61.7	20.6	80.9	0.0	86.2	17.1	100.0	60.9	79.7	0.0

Fig. 6. Result of object detection experiment (Experiment I) using the testing data.

3.2 Experiment II: Comparison with Conventional Machine Learning

Experiment II was executed for comparing the detection accuracy result between deep learning with transfer learning and conventional machine learning (CML). The model used for deep learning with transfer learning is named II-A and that for CML is named II-CML in Experiment II.

Implementation of CML and Numbers of Images. In this research, object detection by HOG features and SVM was executed using Dlib [22], which is a machine learning library. The reason for using Dlib is that objects can be detected with the same creation method of the correct labels as SSD. In object detection using Dlib, if the aspect ratio

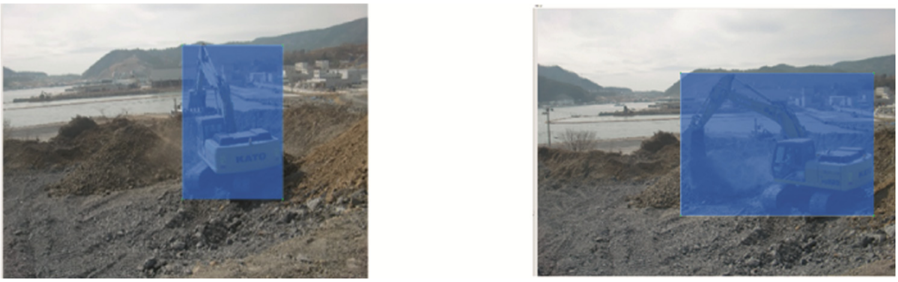


Fig. 7. Sample case of very different aspect ratio of the bounding boxes created for representing correct labels.

of the bounding box created as a correct label is greatly different as shown in Fig. 7, it cannot be used for learning.

Comparison of Detection Results between Deep Learning and CML. Samples of the result of image detection are shown in Fig. 8. In Fig. 8, positive detection by II-A and negative detection by II-CML are shown. In II-CML, Dlib with HOG and SVM was used for detection whereas, in II-A, SSD was used for detection. Figure 9 shows the comparison of the accuracy between II-A and II-CML. The accuracy of the method using deep learning with transfer learning (II-A) is about 90% or higher for all objects while that of II-CML ranges between 30% and 75%. Thus, Experiment shows that the proposed method using deep learning with transfer learning has much higher accuracy compared to CML using HOG and SVM.

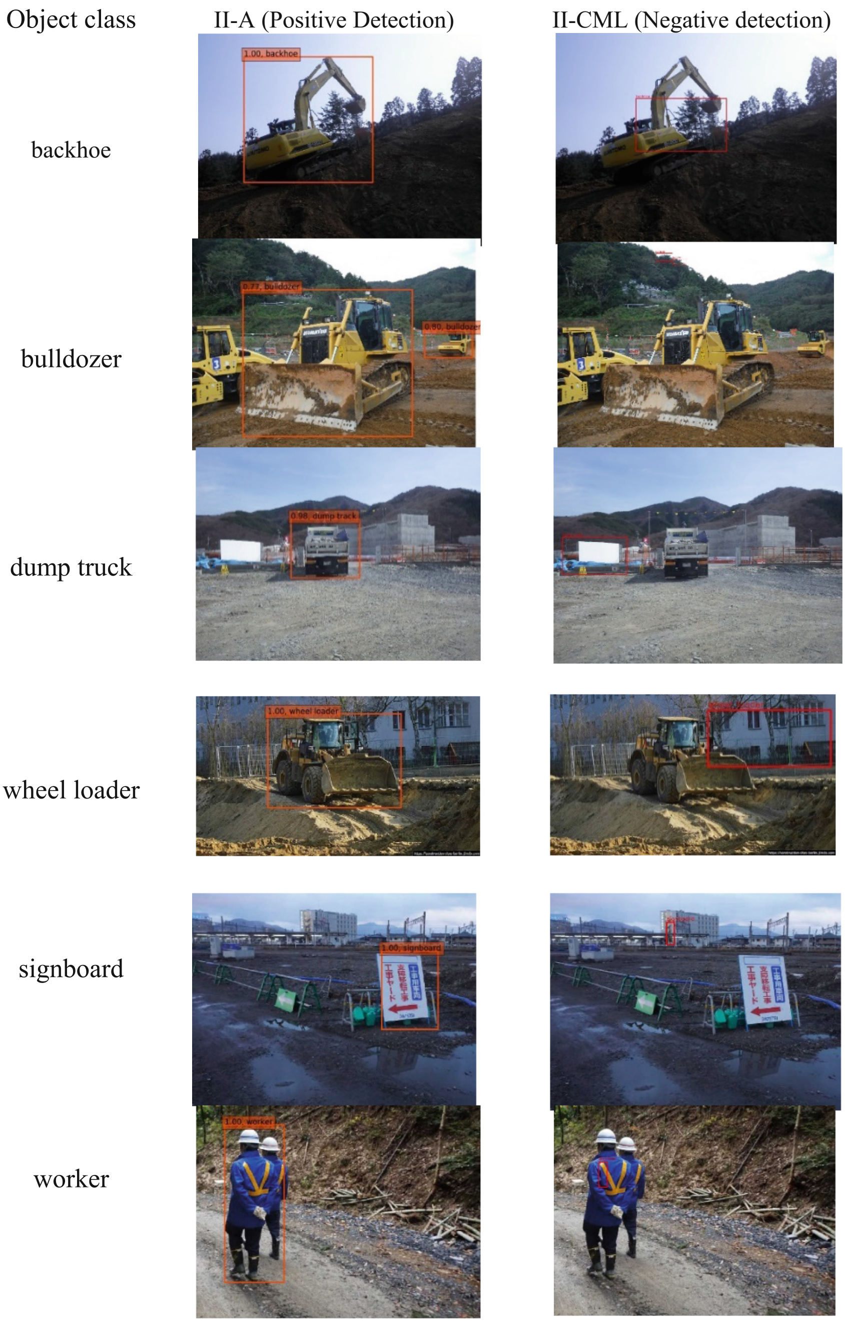
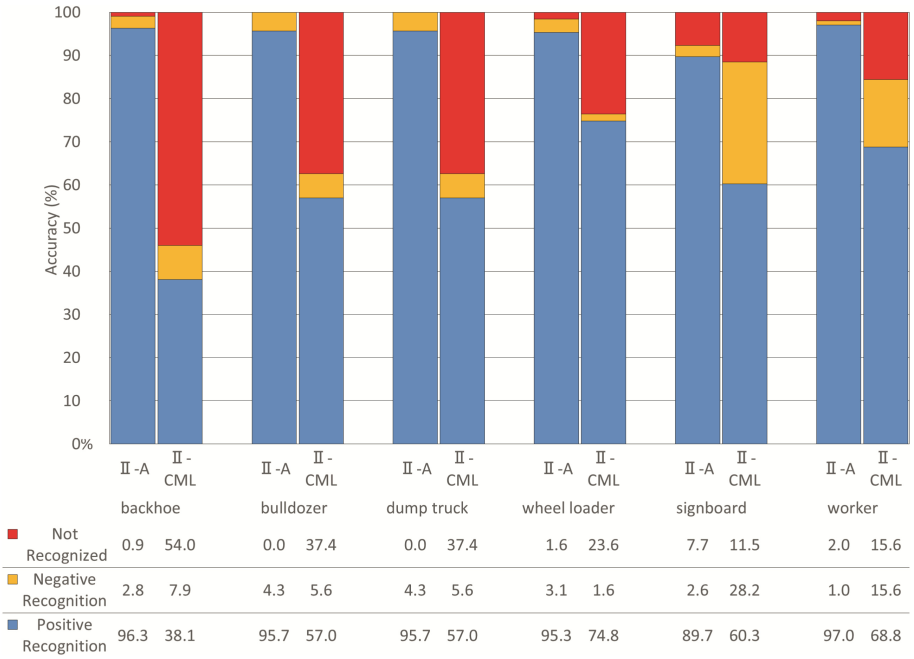


Fig. 8. Sample cases of positive detection by deep learning with transfer learning (II-A) and negative detection by conventional machine learning (II-CML).



	backhoe		bulldozer		dump truck		wheel loader		signboard		worker	
	Ⅱ -A	Ⅱ - CML	Ⅱ -A	Ⅱ - CML	Ⅱ -A	Ⅱ - CML	Ⅱ -A	Ⅱ - CML	Ⅱ -A	Ⅱ - CML	Ⅱ -A	Ⅱ - CML
🟥 Not Recognized	0.9	54.0	0.0	37.4	0.0	37.4	1.6	23.6	7.7	11.5	2.0	15.6
🟨 Negative Recognition	2.8	7.9	4.3	5.6	4.3	5.6	3.1	1.6	2.6	28.2	1.0	15.6
🟦 Positive Recognition	96.3	38.1	95.7	57.0	95.7	57.0	95.3	74.8	89.7	60.3	97.0	68.8

Fig. 9. Testing result of object position detection comparing among deep learning with transfer learning (II-A) and conventional machine learning (II-CML).

4 Conclusions

In this research, a problem was identified in tasks such as collecting, sorting, annotating, storing, deleting, distributing enormous number of digital photographs taken by engineers at construction sites and disaster areas. It was the challenge to manually classify a huge number of photographs. In order to automate these error-prone and cumbersome tasks, an object detection method was proposed, which can detect not just the object class but the position as well as a bounding box. In the proposed method, deep learning (CNNs) with transfer learning was employed because construction-specific objects such as construction machinery, workers, signboards, are not available in the large amount of datasets provided for AI researchers. Most of the datasets usually include ordinary things. The reason for adopting deep learning with transfer learning is that deep learning can automatically determine features from digital images while conventional machine learning requires manual feature selection. The second reason is while deep learning has this advantage, it requires a huge amount of training data, which is difficult to obtain for construction-specific objects that are not available in popular, open datasets. Thus, transfer learning, where a knowledge obtained for one problem can be applied to different but related problems. After the object position detection is done, digital photo files can be classified, copied and put into the designated folders automatically.

Based on the proposed methodology, a system was developed employing SSD and VGG-16. A number of digital images of backhoes, bulldozers, dump trucks, wheel loaders, construction workers, and construction signboards were collected and used for learning. To verify the proposed method, Experiment I, which compares the object detection accuracy between deep learning with and without transfer learning, was executed. As a result, proposed method showed about 80% accuracy while the latter showed very poor performance. Next, Experiment II, which compares the deep learning with transfer learning and conventional machine learning (CML) using HOG features with SVM. The result showed that the proposed method showed about 90% accuracy while that of CML ranged from 30 to 75%.

In conclusion, the proposed method of deep learning with transfer learning can be a useful and effective way to detect object positions from digital images, especially in the area where is domain is specific such as construction sites and huge datasets are not available.

Future work includes developing an object shape detection methodology and apply it to construction-specific objects and improving the object detection accuracy by enhancing the quality of data and increasing the amount of data. As the potential applications of this research in construction and disaster management are more widespread, the authors are planning to apply the future research outcomes to various applications.

References

1. About people in photos on your iPhone, iPad, or iPod touch. https://support.apple.com/en-us/HT207103. Accessed 19 Jan 2018
2. Google products. https://www.google.com/about/products/. Accessed 19 Jan 2018
3. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, pp. 511–518. IEEE (2001)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, vol. 1, pp. 511–518. IEEE (2005)
5. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington D.C., vol. 2, pp. 53–60. IEEE (2004)
6. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In: Tenth IEEE International Conference on Computer Vision, Beijing, vol. 1, pp. 90–97. IEEE (2005)
7. Mitsui, T., Fujiyoshi, H.: Object detection by joint features based on two-stage boosting. In: Proceedings of 2009 IEEE 12th International Conference on Computer Vision Workshops, Kyoto, pp. 1169–1176. IEEE (2009)
8. Simonyan, K., Andrew, Z.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations 2015, San Diego (2015). Preprint CoRR: arXiv:1409.1556
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Ohio, pp. 580–587. IEEE (2014)

10. Girshick, R.: Fast R-CNN. In: Proceedings of 15th IEEE International Conference on Computer Vision, Santiago, pp. 1440–1448. IEEE (2015)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems, Montreal, pp. 91–99. NIPS (2015)
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, pp. 779–788. IEEE (2016)
14. The PASCAL Visual Object Classes Homepage. http://host.robots.ox.ac.uk/pascal/VOC/. Accessed 19 Jan 2018
15. Caltech 101. http://www.vision.caltech.edu/Image_Datasets/Caltech101/. Accessed 19 Jan 2018
16. COCO dataset. http://cocodataset.org/. Accessed 19 Jan 2018
17. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
18. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/about.html. Accessed 19 Jan 2018
19. Transfer Learning – Machine Learning's Next Frontier. http://ruder.io/transfer-learning/index.html. Accessed 19 Jan 2018
20. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
21. LabelImg. https://github.com/tzutalin/labelImg. Accessed 1 Sept 2017
22. Dlib C++ Library Python API. http://dlib.net/python/index.html. Accessed 1 Sept 2017

Data Driven Analytics (Machine Learning) for System Characterization, Diagnostics and Control Optimization

Jinkyoo Park[1], Max Ferguson[2], and Kincho H. Law[2](✉)

[1] Korea Advanced Institute of Science and Technology (KAIST),
Daejeon, Korea
[2] Stanford University, Stanford, CA 94305, USA
law@stanford.edu

Abstract. This presentation discusses the potential use of machine learning techniques to build data-driven models to characterize an engineering system for performance assessment, diagnostic analysis and control optimization. Focusing on the Gaussian Process modeling approach, engineering applications on constructing predictive models for energy consumption analysis and tool condition monitoring of a milling machine tool are presented. Furthermore, a cooperative control optimization approach for maximizing wind farm power production by combining Gaussian Process modeling with Bayesian Optimization is discussed.

Keywords: Machine learning · Predictive models · Data-driven optimization

1 Introduction

The last decade has seen an increasing number of research and applications of machine learning [1]. As sensor technologies, data acquisition systems, and data analytics continue to improve, companies can now effectively and efficiently collect large and diverse volumes of data and gain valuable insights from the data. There have been a growing interest in applying machine learning to draw insights gained from the data in engineering [2]. These data-driven approaches are able to find highly complex and nonlinear patterns in data of different types and transform the raw data into useful models. Machine learning techniques has applied for prediction [3, 4], detection and classification [5, 6], forecasting [7] and many problems of specific interests.

Engineering systems are inherently dynamic, uncertain, and complex. Majority of machine learning techniques provide point predictions that do not usually consider uncertainties in the models. As prediction and forecasting of future consequences carry many unknowns and uncertainties, a prediction model should provide some quantification of uncertainty for informed decision making [8–10]. In this paper, we discuss methods that are based on Bayesian statistic inference and illustrate their potential applications for performance characterization, condition diagnostic and control optimization of physical systems, for which analytical modeling of such systems can be difficult.

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 16–36, 2018.
https://doi.org/10.1007/978-3-319-91635-4_2

Gaussian Process Regression (GPR) has received increased attention in the machine learning community and has been applied to a variety of applications [8, 11–16]. Without predefining the basis functions, the nonparametric GPR can be used to approximate a target function that represents the complex input and output relationships based on the observed data and predicts a target output with quantified uncertainty. This paper discusses the potential applications of GPR for system characterization and diagnostics, using a manufacturing milling machine tool as an illustrative physical system.

Machine learning has been broaden from building predictive and forecasting models to supporting optimization and decision making problems [17]. For many complex physical systems, constructing the analytical model and the objective function for optimal decision making can be difficult. An alternative would be to establish a model using the data collected from the physical system. A data-driven approach would involve using the sampled data to construct the model while searching for the optimal value that maximizes the objective target function at the same time. For implementation in a physical environment, the number of sampled trials should be kept to a minimum as each trial would likely involve the execution of some physical actions. Using Gaussian Process (GP) as a way to approximate the target function, Bayesian Optimization (BO) has been shown effective in finding the optimal values of a target function with a small number of sampled trials [18]. For efficient sampling, BO uses an acquisition function that takes advantage of the estimated probability distribution of the target function so that the number of function evaluations (i.e. executions of physical actions) is kept small. In other words, BO iteratively approximates the input and the output relationship of a target system using Gaussian Process (GP) regression and utilizes an acquisition function with the learned model to determine the trial inputs that can potentially improve the target values [18–20]. Bayesian optimization has been applied to many optimization problems such as the multi-armed bandit and sensor placements [21–24]. When implementing with a physical system, the trial actions are typically constrained by the physical limitation of the system. Furthermore, it may not be desirable to impose abrupt changes in successive actions. We propose a Bayesian Ascent (BA) algorithm which augments BO with trust region constraints to ensure that successive actions do not deviate significantly [25]. This paper discusses an application of the Bayesian Ascent algorithm for maximizing the total power production of a wind farm with multiple wind turbines as an illustrative implementation of the algorithm on a physical system.

The purpose of this paper is intended primarily to illustrate the potential applications of Bayesian-based machine learning approach by reviewing a number of example implementations of the approach in engineering systems. The paper is organized as follows: Section 2 provides a brief description of the Gaussian Process Regression method with illustrative examples involving the development of an energy prediction model and a tool condition diagnostic model for a milling machine tool. Section 3 discusses Bayesian Optimization and its adoption for the wind farm power production problem. Section 4 summarizes the paper with a brief discussion.

2 System Characterization and Diagnosis with GPR

This section discusses the use of input and output data of a physical system to characterize the performance of the system and to perform diagnostic analysis. Specifically, Gaussian Process Regression (GPR) is employed to build predictive models for energy performance and tool conditions of a milling machine tool. Prior to the training process, the raw data collected are pre-processed to extract features that are deemed useful for the construction and execution of the predictive models. Using the feature data as training data points, GPR then approximates a target function that represents the input and output relationships without predefining a set of basis functions. Given a new input, the approximated target function is then used to predict the target output with uncertainty quantification. One desirable property of a GPR model is its ability to capture complex input and output relationships with a small number of hyper-parameters. Moreover, a GPR model can be trained with relatively small number of experimental training data sets but is able to give reasonable predictive estimation quantified with uncertainty measures.

2.1 Gaussian Process Regression (GPR)

Detailed description of Gaussian Process Regression (GPR) can be found in the literature (for examples, see [11, 12]). Given a data set $\boldsymbol{D}^n = \{(\boldsymbol{x}^i, y^i) | i = 1, \ldots, n\}$ with n samples, where $\boldsymbol{x}^{1:n} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)$ and $\boldsymbol{y}^{1:n} = (y^1, \ldots, y^n)$ denote, respectively, the inputs and the corresponding (possibly noisy) output observations of the target function values $\boldsymbol{f}^{1:n} = (f(\boldsymbol{x}^1), \ldots, f(\boldsymbol{x}^n))$, GPR constructs a posterior distribution $p(f^{new}|\boldsymbol{D}^n)$ on the function value $f^{new} = f(\boldsymbol{x}^{new})$ corresponding to an unseen input $\boldsymbol{x}^{new}$. Each observed (or response) value is assumed to contain some random noise ϵ^i, such that $y^i = f(\boldsymbol{x}^i) + \epsilon^i$, where $\epsilon^i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ is assumed to be independent and identically distributed Gaussian with noise variance σ_ϵ^2.

For Gaussian Process (GP), the prior on the function values $p\left(\boldsymbol{f}^{1:n}\right)$ can be fully described by its mean function $m(\boldsymbol{x}) = \mathrm{E}[f(\boldsymbol{x})]$ and a covariance (kernel) function $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$:

$$p\left(\boldsymbol{f}^{1:n}\right) = GP\left(m(\boldsymbol{x}), k\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right)\right) \tag{1}$$

The kernel function $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$ represents a geometrical distance measure that the more closely located inputs would be more correlated in terms of their function values $f(\boldsymbol{x}^i)$ and $f(\boldsymbol{x}^j)$.

A variety of kernel functions have been proposed and described in the literature [12, 26]. One common choice for a GPR model is the stationary and (infinitely) differentiable *squared exponential* (SE) kernel:

$$k_{SE}\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right) = \sigma^2 \exp\left(-\frac{1}{2\lambda^2} \left\|\boldsymbol{x}^i - \boldsymbol{x}^j\right\|^2\right), \tag{2}$$

where the kernel function is described by two hyperparameters, namely the signal variance σ^2 and the length scale λ. While the SE kernel function is a good choice for many applications for a feature vector with multiple variables (or dimensions), it does not allow the length scale to vary for each dimension in the feature vector. An alternative is to use an *automatic relevance determination* (ARD) kernel, which assigns a different length scale to each dimension. The ARD squared exponential (ARD-SE) kernel, which is essentially a product of the SE kernels over different dimensions, can be expressed as:

$$k_{ARD-SE}\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right) = \sigma^2 \exp\left(-\frac{1}{2}\left(\boldsymbol{x}^i - \boldsymbol{x}^j\right)^T \operatorname{diag}(\boldsymbol{\lambda})^{-2}\left(\boldsymbol{x}^i - \boldsymbol{x}^j\right)\right). \tag{3}$$

where the parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_i, \ldots, \lambda_m)$ is referred to as the characteristic length scales that quantify the relevancy of the input features in $\boldsymbol{x^i} = \left(x_1^i, \ldots x_k^i, \ldots, x_m^i\right)$ with m variables. An ARD-SE kernel provides the flexibility to adjust the relevance (weight) of each parameter in the feature vector. A large length scale λ_i indicates weak relevance for the corresponding input feature $\boldsymbol{x}^i$ and vice versa. Other kernel functions such as linear, periodic, Matern kernels exist [12, 26, 27]. A kernel function can also be constructed by combining a number of kernel functions via, for examples, addition and multiplication of the kernel functions [12, 26]. The kernel function produces a positive semi-definite symmetric kernel matrix $\mathbf{K}$ whose (i,j)th entry is $\mathbf{K}_{ij} = k(\boldsymbol{x}^i, \boldsymbol{x}^j)$. The type of kernel function chosen is problem dependent, and can strongly affect the representability of the GPR model and influence the accuracy of the predictions.

Including the noise model, the covariance function is parameterized (i.e., defined) by the hyperparameters jointly denoted by $\boldsymbol{\theta} = (\sigma_\epsilon, \sigma, \boldsymbol{\lambda})$. The regression model can be trained by finding the hyperparameters $\boldsymbol{\theta} = (\sigma_\epsilon, \sigma, \boldsymbol{\lambda})$ that maximizes the marginal log-likelihood of the training data $\boldsymbol{D}^n = \{(\boldsymbol{x}^i, y^i)|i = 1, \ldots, n\}$ as [11, 12]:

$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p\left(\boldsymbol{y}^{1:n}|\boldsymbol{\theta}\right) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left(-\frac{1}{2}\left(\boldsymbol{y}^{1:n}\right)^T\left(\mathbf{K} + \sigma_\epsilon^2\mathbf{I}\right)^{-1}\boldsymbol{y}^{1:n} - \frac{1}{2}\log\left|\mathbf{K} + \sigma_\epsilon^2\mathbf{I}\right| - \frac{n}{2}\log 2\pi\right)\end{aligned} \tag{4}$$

As long as the kernel function is differentiable with respect to its hyper-parameters $\boldsymbol{\theta}$, various off-the-shelf mathematical programming tools, such as GPML [28], scikit-learn [29] and others, can be employed for the optimization problem.

Once the optimized hyperparameters are obtained, the trained GPR model is fully characterized. Denoting a newly observed input $\boldsymbol{x}^{new}$:

$$\boldsymbol{x}^{new} = \left(x_1^{new}, \ldots x_k^{new}, \ldots x_m^{new}\right)$$

The function value $f(\boldsymbol{x}^{new}) \sim N\left(\mu(\boldsymbol{x}^{new}|\boldsymbol{D}^n), \sigma^2(\boldsymbol{x}^{new}|\boldsymbol{D}^n)\right)$ can be described with the mean and variance functions expressed, respectively, as [11, 12]:

$$\mu(\boldsymbol{x}^{new}|\boldsymbol{D}^n) = \boldsymbol{k}^T\left(\mathbf{K} + \sigma_\epsilon^2\mathbf{I}\right)^{-1}\boldsymbol{y}^{1:n} \tag{5}$$

$$\sigma^2(\boldsymbol{x}^{new}|\boldsymbol{D}^n) = k(\boldsymbol{x}^{new}, \boldsymbol{x}^{new}) - \boldsymbol{k}^T\left(\mathbf{K} + \sigma_\epsilon^2\mathbf{I}\right)^{-1}\boldsymbol{k} \tag{6}$$

where $\boldsymbol{k}^T = (k(\boldsymbol{x}^1, \boldsymbol{x}^{new}), \ldots, k(\boldsymbol{x}^n, \boldsymbol{x}^{new}))$. That is, $\mu(\boldsymbol{x}^{new}|\boldsymbol{D}^n)$ and $\sigma^2(\boldsymbol{x}^{new}|\boldsymbol{D}^n)$ can be used as the scoring functions for predicting the hidden function output $f(\boldsymbol{x}^{new})$ corresponding to the input data $\boldsymbol{x}^{new}$.

2.2 Characterizing Energy Consumption of a Milling Machine

Monitoring and optimizing energy efficiency of manufacturing processes has become a priority in the manufacturing industry. This section discusses how a GPR-based energy prediction model for a milling machine is established. An energy prediction model can provide a better understanding of how different operational strategies may influence the energy consumption pattern of a machine tool and enable selection of optimal strategy with efficient operations for machining a part.

Figure 1 shows the basic set up of a Mori Seiki NVD 1500DCG 3-axis milling machine tool. The machining process data, such as process parameters, NC blocks, and tool positions, are collected from the FANUC controller and the power time series data is collected using a High Speed Power Meter (HSPM). With recent technologies and standards, such as MTConnect [30], it is now possible to track variations in energy consumption by different machine operations [31]. The raw data collected includes the timestamp, time series data for power consumption, feed rate and spindle speed, and numerical control (NC) code information. The raw data is then post-processed to derive process parameters such as average feed rate, average spindle speed, cumulative energy consumption, volume of material removed, depth of cut and cutting strategy for each NC code block. As our interest is to directly relate machine operations to energy consumption, the process parameters are selected as the input features that include:

- Feed rate, x_1: the velocity at which the tool is fed
- Spindle speed, x_2: rotational speed of the tool
- Depth of cut, x_3 : depth of material that the tool is removing
- Active cutting direction, x_4: 1 for x-axis, 2 for y-axis, 3 for z-axis, and 4 for x-y axes

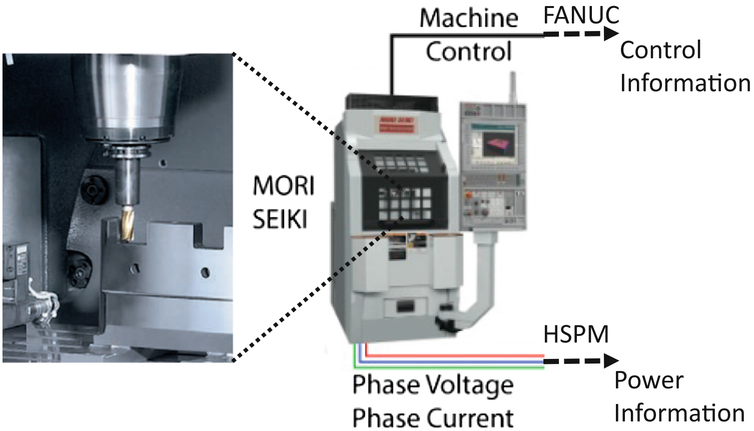


Fig. 1. A Mori Seiki NVD 1500DCG 3-axis milling machine

- Cutting strategy, x_5: the method for removing material with 1 for conventional cut, 2 for climbing cut, and 3 for a combination of both.

Each operation, i.e. a feature data, corresponds to a single NC block. The energy consumption, E, for each NC block is obtained by numerically integrating the power time series recorded by the HSPM over the duration of the NC block. Additionally, the length of the tool path, l, in a single NC block is computed using the lengths of cut in the x-, y- and z-directions. The output parameter, y, is defined as the energy consumption per unit length (i.e. $y^i = E^i/l^i$ for the i^{th} NC block operation). The hardware platform, the data acquisition system, the experimental design, and the data processing techniques have been described in details by Bhinge et al. [32].

For training and testing, a total of 18 parts were machined and created 3,214 (cleansed) data sets (NC code blocks). Once the mean energy density function $\mu(\boldsymbol{x}^i|\boldsymbol{D}^{\boldsymbol{train}})$ and associated standard deviation function $\sigma(\boldsymbol{x}^i|\boldsymbol{D}^{\boldsymbol{train}})$ of the GPR model, we can apply the model to estimate the energy consumption $\hat{E}^i$ and the corresponding standard deviation S^i on the estimated energy consumption value as:

$$\hat{E}^i = \mu\left(\boldsymbol{x}^i|\boldsymbol{D}^{\boldsymbol{train}}\right) \times l_i \tag{7}$$

$$S^i = \sigma\left(\boldsymbol{x}^i|\boldsymbol{D}^{\boldsymbol{train}}\right) \times l_i \tag{8}$$

Details on developing the GPR model have been reported in [13, 14]. Even with relatively small number of data sets, the predictive results are highly accurate, in that the relative errors (i.e. the difference between the measured and estimated energy consumption with respect to the measure value) for blind tests (of 3 new parts) are less than 6%. Furthermore, the standard deviation of the predictions is within 5% of the measured values.

This study illustrates that the GPR models can be established to represent the complex relationship between the input machining parameters and output energy consumption, and construct a prediction function for the energy consumption with confidence bounds. Furthermore, since the input features are directly related to the NC code information (which can often be generated from a CAD model), the saved GPR model can be used to predict energy consumption when machining a new part as well as to select the best tool path for machining a part with minimal energy [13, 14]. Lastly, as shown in Fig. 2, the predictive energy model can be employed together with the load measurements for real time monitoring of machine tool operations.

2.3 Diagnosis of Tool Conditions

This section describes an application of GPR in developing a predictive model for tool condition diagnostic. Researchers have previously demonstrated that the condition of a machine tool can be inferred from features of vibration and audio time series [33–35]. With the availability of low cost sensors, it is possible to collect real time vibration and audio data from critical locations inside a manufacturing machine tool.

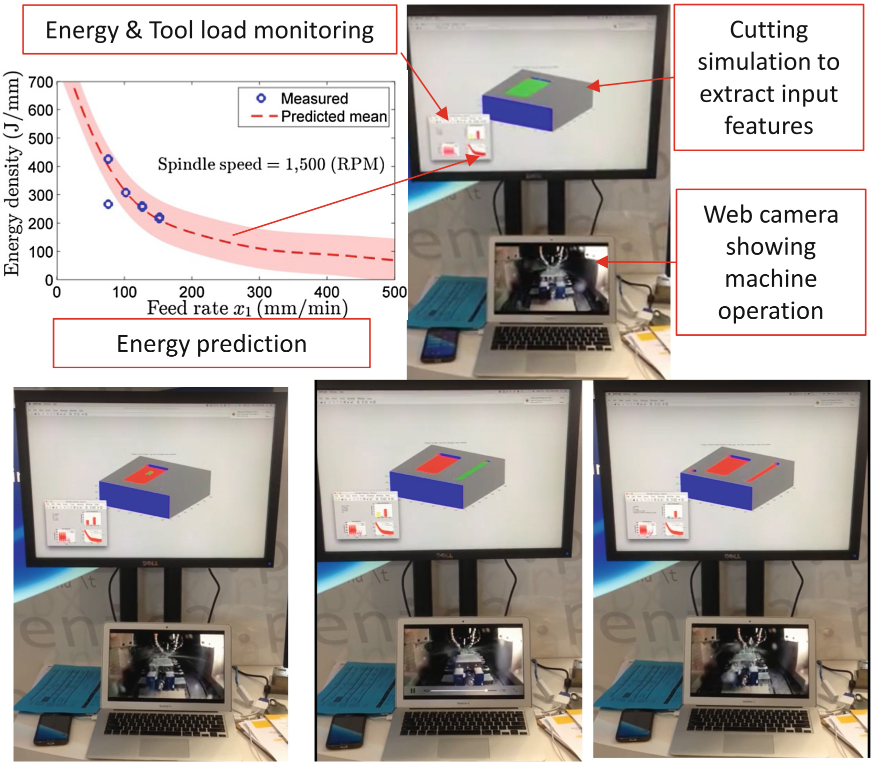
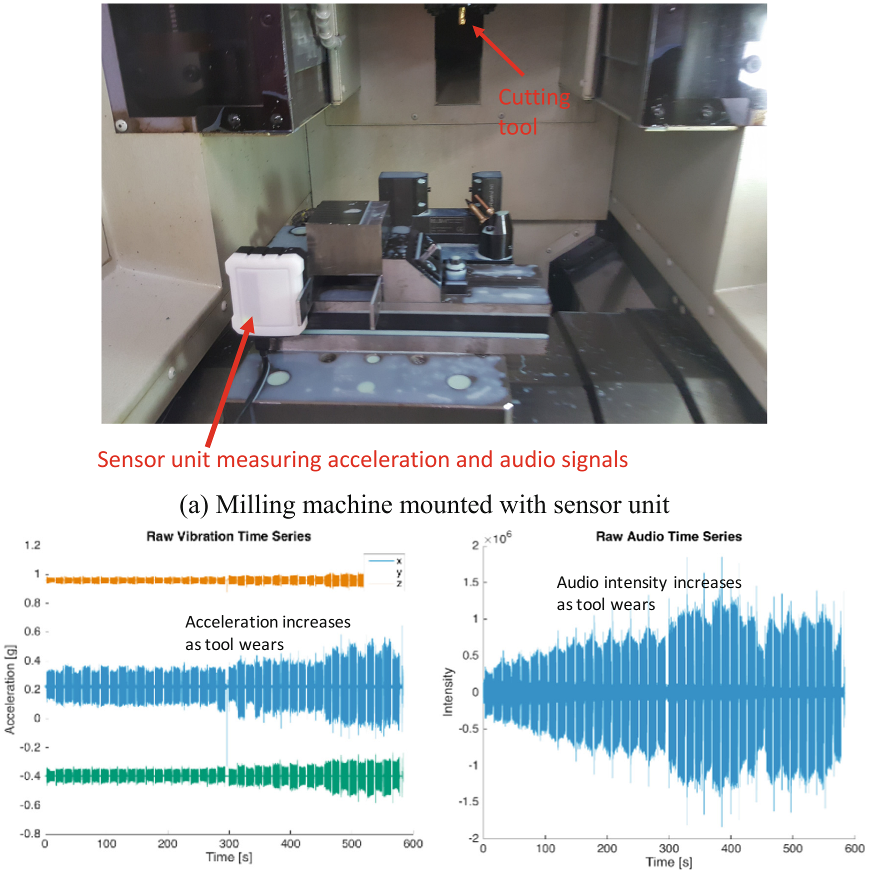

Energy & Tool load monitoring
Cutting simulation to extract input features
Web camera showing machine operation
Energy prediction
Energy density (J/mm)
700
600
500
400
300
200
100
0
0
100
200
300
400
500
Feed rate x_1 (mm/min)
Measured
Predicted mean
Spindle speed = 1,500 (RPM)

Fig. 2. Continuous monitoring of machine tool operations with energy prediction

As shown in Fig. 3(a), a waterproof sensor unit from Infinite Uptime, Inc. is attached to the vise of the Mori Seiki NVD 1500DCG 3-axis milling machine. The sensor unit is capable of measuring both the audio and triaxial acceleration signals inside the milling machine. In the experimental set up, the acceleration signal is recorded in the *x*-, *y*- and *z*-direction with a sampling rate of 1000 Hz to capture the 200 Hz signal generated by the cutting tool when the spindle rate is set to 3000 RPM. The audio signal is recorded at 8000 Hz. Data is streamed from the sensor to a laptop computer using a Universal Serial Bus (USB) connection. Figure 3(b) shows the recorded time series data. As can be seen in the figure, vibration and audio signals tend to increase as the tool deteriorates.

The milling machine was programmed to produce a number of simple parts by removing material from a solid steel block until the cutting tool became severely damaged, or the cutting tool broke. Tests were conducted using an Atrax solid carbide 4-flute square end mill, and a continuous supply of coolant. As described previously, the machining data, such as tool position and rotation speed, can be recorded from the FANUC controller and streamed to a laptop computer, along with a timestamp. The data is then post-processed to extract the behavior of the machine from the raw numerical control (NC) code [32].



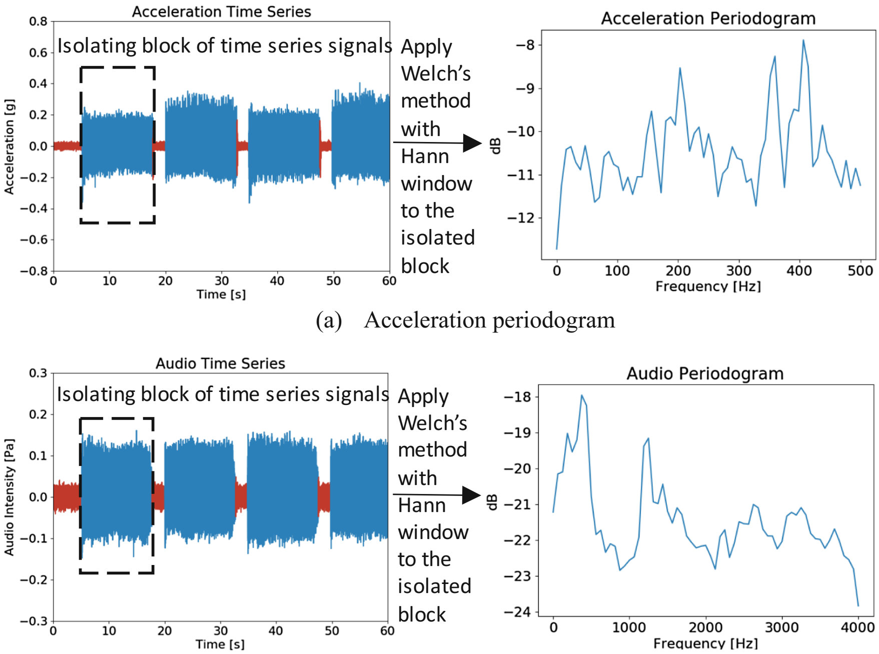
(a) Milling machine mounted with sensor unit

(b) Recorded vibration and acoustic time series signals

Fig. 3. Experimental set up for tool condition diagnosis

In this work, we define the condition of the milling machine tool $y \in [0, 1]$, based on the remaining lifetime of the tool, as estimated after manually examining the tool with a microscope. The scale is defined such that 1 indicates a new tool in perfect condition, and 0.5 indicates the condition at which the tool would need to be replaced in an industrial setting as judged by a machine operator.

To extract the relevant features for training the model, the periodic audio and vibration time series signals are transformed into power spectra [36–38]. We employ Welch's method with a Hann window function to reduce noise and spectral leakage when generating the power spectra [37]. Figure 4 shows the data processing (feature extraction) step. A discrete time series signal s, is first divided into K successive blocks s_m, using a window function w:



(a) Acceleration periodogram

(b) Audio periodogram

Fig. 4. Transforming time series data to periodograms as features

$$s_m(n) = w(n)s(n+mR), \quad n = 0\ldots M-1, \quad m = 0\ldots K-1 \tag{9}$$

where M is the length of the window, R is the window hop size and the Hann window function $w(n)$ employed is given as [37]:

$$w(n) = \begin{cases} 0.5\left(1-\cos\left(\frac{2\pi n}{M-1}\right)\right) & \textit{if } n \leq M-1; \\ 0 & \textit{otherwise}. \end{cases} \tag{10}$$

The periodogram of the m^{th} block is then calculated using the Fourier transform:

$$\boldsymbol{p}_m(\omega_k) = \frac{1}{M}\left|\sum_{n=0}^{N-1} s_m(n)e^{-\frac{i2\pi nk}{N}}\right|^2. \tag{11}$$

where ω_k is the k^{th} point in the discretized frequency domain. The Welch estimate of the power spectral density is given by:

$$\hat{\boldsymbol{s}}(\omega_k) = \frac{1}{K}\sum_{m=0}^{K-1}\boldsymbol{p}_m(\omega_k). \tag{12}$$

Each periodogram is obtained from a windowed segment of the time series. It may be worth noting that the data processing and feature extraction step can be implemented in real time (simultaneously with the model training and scoring procedures). In our study, a window overlap of 50% is used and a Hann window length of 256 points is chosen for both the vibration and audio signals.

In addition to the vibration periodogram $\hat{\boldsymbol{s}}_v^i \in \mathbb{R}^{256}$ and acoustic periodogram $\hat{\boldsymbol{s}}_a^i \in \mathbb{R}^{256}$ for each milling machine operation i, the previous condition c^i of the cutting tool is also included in the feature vector since the cutting tool tends to deteriorate gradually. Altogether, the input feature vector for each milling operation i is denoted as:

$$\boldsymbol{x}^i = \begin{bmatrix} c^i \\ \hat{\boldsymbol{s}}_v^i \\ \hat{\boldsymbol{s}}_a^i \end{bmatrix}. \tag{13}$$

During the training procedure, the previous tool condition is known, so c^i can be assigned as the previously measured condition y^{i-1} :

$$c_{training}^i = \begin{cases} 1 & \text{for } i = 1, \\ y^{i-1} & \text{otherwise.} \end{cases} \tag{14}$$

However, when making a new prediction using the scoring procedure, while the vibration and audio periodograms can be calculated immediately after each operation is performed by the milling machine, the previous tool condition value y^{i-1}, will not be known. Instead, we use the previous tool condition prediction, $\hat{y}^{i-1}$, in the scoring procedure, that is:

$$c_{scoring}^i = \begin{cases} 1 & \text{for } i = 1, \\ \hat{y}^{i-1} & \text{otherwise.} \end{cases} \tag{15}$$

As a result, the prediction process step is computationally a recursive process.

When selecting a kernel function for a physical model, it is important to carefully define a kernel that can capture the (relevant) relationship among the feature data with a small number of hyperparameters (as to reduce the problem dimensionality and computations). When using the ARD squared exponential kernel, the number of hyperparameters grows linearly with the dimensional size of the feature vector, making the model prone to overfitting [39]. Instead, we choose to define a composite *sum of square exponential* (SSE) kernel by combining the SE kernels for each data type [40]:

$$\begin{aligned} k_{SSE}\left(\boldsymbol{x}^i, \boldsymbol{x}^j\right) = & \sigma_1^2 \exp\left(-\frac{1}{2\lambda_1^2}\left\|c^i - c^j\right\|^2\right) + \sigma_2^2 \exp\left(-\frac{1}{2\lambda_2^2}\left\|\hat{\boldsymbol{s}}_v^i - \hat{\boldsymbol{s}}_v^j\right\|^2\right) \\ & + \sigma_3^2 \exp\left(-\frac{1}{2\lambda_3^2}\left\|\hat{\boldsymbol{s}}_{\boldsymbol{a}}^i - \hat{\boldsymbol{s}}_{\boldsymbol{a}}^j\right\|^2\right), \end{aligned} \tag{16}$$

where $\sigma_1, \sigma_2, \sigma_3, \lambda_1, \lambda_2$ and λ_3 are the parameters to be determined for the SSE kernel function. Including the noise term σ_ϵ^2, the SSE kernel function consists of seven hyperparameters, $\boldsymbol{\theta} = (\sigma_1, \sigma_2, \sigma_3, \lambda_1, \lambda_2, \lambda_3, \sigma_\epsilon)$. The SSE kernel has significantly fewer hyperparameters than the ARD-SE kernel, but still allows the length scale of the previous state and the periodograms to be adjusted independently.

We randomly select 14 experiments for the training set and 4 experiments for the testing set. In the training process, the models for climb-cutting and conventional-cutting are built separately. The two trained models are then used together to predict the condition of the tool for each test case. As illustrated in Fig. 5 which shows a typical result on the tool, even with a small number of training data sets, the composite kernel performs relatively well for predicting the condition of the cutting tool, particularly within the lightly damaged region (i.e. tool condition > 60–70%) when the tool is considered operational. For worn tool (say, tool condition < 50%), the confidence level becomes high since there are less training data points because the experimental tests often need to be stopped abruptly before or when the tool became severely worn or broken. In short, this study illustrates the potential use of GPR as a means for predictive tool condition monitoring. It is worth mentioning that the methodology is designed for real time automated operation, from data acquisition to model training and model prediction.

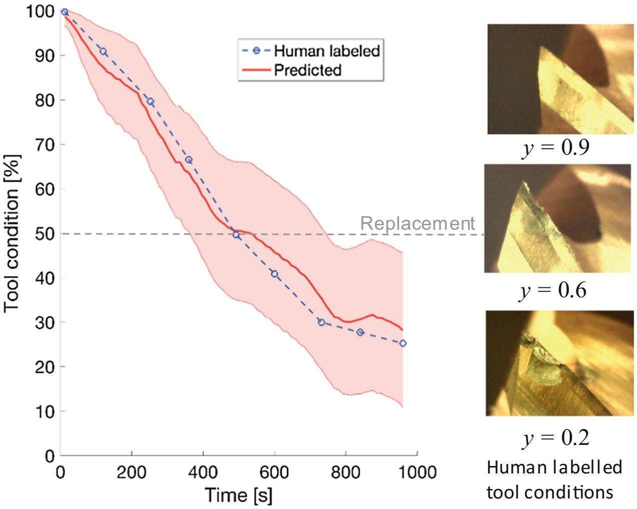


Fig. 5. Illustration of tool condition prediction result. The shaded region represents the 90% confidence interval for each prediction.

3 System Control Optimization

This section discusses real time control of a physical system, for which the construction of the analytical model and the objective function is difficult. The strategy is to iteratively select a series of trial input actions that potentially optimize the objective, observe the corresponding outputs, and learn about the unknown target function based on the inputs and the outputs. One important consideration for control of a physical system using a data-driven approach is to ensure that the number of trials is kept small as a trial would involve execution of corresponding control actions of the physical system. Utilizing GPR to establish a learned model based on the input trials and output measurements, Bayesian Optimization (BO) has been found effective in optimizing a target function using a small number of trials. Furthermore, a trust region constraint is imposed to avoid abrupt changes of the control actions. To illustrate, we discuss an application of BO to maximize the power production of a wind farm with multiple wind turbines in an experimental wind tunnel study.

3.1 Bayesian Optimization

Bayesian Optimization (BO) iteratively approximates the input and the output relationship of a target system using Gaussian Process (GP) regression and uses the learned model to determine the trial inputs that can potentially improve the target values [18–20]. Following the GPR procedure described in the previous section, denote $\boldsymbol{x} = (x_1, \ldots, x_m)$ as the m input control actions and let $f(\boldsymbol{x})$ be the unknown target function for the output measurement y inferred by $\boldsymbol{x}$. The output measurement $y = f(\boldsymbol{x}) + \epsilon$, is assumed to include noise ϵ which follows a Gaussian distribution, i.e., $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ [18–20]. At the nth iteration, using the historical input-output data $\boldsymbol{D}^n = \left\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^i, y^i), \ldots, (\boldsymbol{x}^n, y^n)\right\}$, where $\boldsymbol{x}^i = \left(x_1^i, \ldots, x_m^i\right)$ and y^i represent, respectively, the input actions and the corresponding output measurement at the ith iteration, BO attempts to improve the target function by executing three basic steps:

- Learning: model the unknown target function $f(\boldsymbol{x})$ using Gaussian Process (GP) regression.
- Optimization: select the next trial input $\boldsymbol{x}^{n+1}$ that is useful for increasing (exploitation) and learning (exploration) the objective function $f(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}|\boldsymbol{D}^n), \sigma^2(\boldsymbol{x}|\boldsymbol{D}^n))$.
- Observation: execute the selected actions $\boldsymbol{x}^{n+1}$ and obtain the corresponding output y^{n+1} from the target system.

The new input and output pair $(\boldsymbol{x}^{n+1}, y^{n+1})$ will then be used to update the regression model for the target function $f(\boldsymbol{x})$ in the next iteration. The key of BO lies in the optimization step for selecting a trial action to improve towards the optimal control actions. The goal is to select the next input $\boldsymbol{x}^{n+1}$ in order to *learn* more about the target function and to *improve* the target value at the same time.

As discussed, the GPR is defined by a mean function $\mu(\boldsymbol{x}|\boldsymbol{D}^n)$ and a variance function $\sigma^2(\boldsymbol{x}|\boldsymbol{D}^n)$. The strategy would be to select the next input by exploiting the current belief about the target system by maximizing the current mean function $\mu(\boldsymbol{x}|\boldsymbol{D}^n)$ as well as exploring the uncertainty (variance) around the selected input by maximizing the variance function $\sigma^2(\boldsymbol{x}|\boldsymbol{D}^n)$. In general, the next sampling input is selected as the one that maximizes an acquisition function that incorporates both the aspects of exploration and exploitation as illustrated as shown in Fig. 6. For implementation with

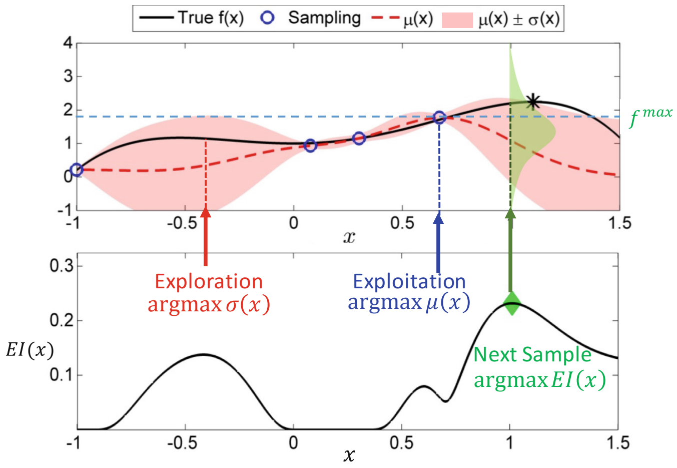


Fig. 6. Acquisition function incorporating exploration and exploitation

a physical system, we select the next input $\boldsymbol{x}^{n+1}$ based on maximizing the expected improvement (EI) acquisition function as [41]:

$$\boldsymbol{x}^{n+1} = \arg \max_{\boldsymbol{x} \in \boldsymbol{A} \cap \boldsymbol{T}} \mathrm{EI}(x) \triangleq \arg \max_{\boldsymbol{x} \in \boldsymbol{A} \cap \boldsymbol{T}} \mathrm{E}[\max\{0, f(\boldsymbol{x}) - f^{max}\} | \boldsymbol{D}^n] \tag{17}$$

Here, $\boldsymbol{A} := \{\boldsymbol{x} | \boldsymbol{x}^l \le \boldsymbol{x} \le \boldsymbol{x}^u\}$ defines the ranges of allowable values for the actions $\boldsymbol{x}$, $\boldsymbol{T} := \{\boldsymbol{x} | \| x_i - x_i^{max} \| < \tau_i \text{ for } i = 1, \ldots, m\}$ is a trust region used to restrict the change of actions, and $f^{max} = \max_{\{\boldsymbol{x} \in \boldsymbol{x}^{1:n}\}} \mu(\boldsymbol{x} | \boldsymbol{D}^n)$ is the maximum target function value estimated in the (current) nth iteration. The ranges for system constraints $\boldsymbol{A}$ define the limits of trial actions allowed by the physical system. For a physical system, abruptly changing the control actions may not be desirable. The trust region $\boldsymbol{T}$ is imposed with a proximity constraint, τ_i, that limits the range of change allowable for the actions per each iterative step. We call the BO with the trust region constraint Bayesian Ascent (BA) method, in that the search follows the ascending direction estimated probabilistically from a sequence of observations [25].

The expected improvement acquisition function, $EI(x)$, as shown in Eq. 17 is chosen in such a way that no calibrated parameters are needed to balance the exploration and exploitation strategies. The quantity $\max\{0, f(\boldsymbol{x}) - f^{max}\}$ denotes the improvement toward the maximum output $f(\boldsymbol{x})$ with respect to the estimated maximum function value f^{max}. Here we use the estimated f^{max} instead of the actually observed maximum response y^{max} because the measurement value y^{max} may have large noise. Given the estimated f^{max}, the value of the expected improvement $\mathrm{EI}(\boldsymbol{x})$ can be analytically derived using the distribution of the target function $f(\boldsymbol{x})$ at $\boldsymbol{x}$ [41]:

$$EI(x) = \begin{cases} (\mu(x) - f^{max})\Phi(Z) + \sigma(x)\phi(Z) & \mathit{if}\ \sigma(x) > 0 \\ 0 & \mathit{if}\ \sigma(x) = 0 \end{cases} \tag{18}$$

where $\phi(\cdot)$and $\Phi(\cdot)$ denote, respectively, the probability and cumulative distribution functions, and $Z = \frac{\mu(x) - f^{max}}{\sigma(x)}$.

In the observation phase, the selected actions $\boldsymbol{x}^{n+1}$ are executed and the corresponding output measurement y^{n+1} is observed. The collected new data point $(\boldsymbol{x}^{n+1}, y^{n+1})$ is then used to update the regression model in the learning phase of the next iteration. Details on the Bayesian Ascent method can be found in [25].

3.2 Cooperative Maximization of Wind Farm Power Production

This section discusses an application of the BA algorithm for finding the optimum coordinated control actions using only the measurement data observed in a physical system. In a wind farm with multiple wind turbines, wakes formed by the upstream wind turbines can decrease the attacking wind speed and, thus, reduce the power production of the downstream wind turbines. Typically, each wind turbine would independently adjust its yaw and blades to maximize its own power production without taking into consideration of the power production of other wind turbines; we refer this non-cooperative, independent actions as a greedy strategy. On the other hand, the cooperative optimization strategy would be to adjust the yaws and the blades of the wind turbines such that the total wind farm power production is maximized. Mathematical programming approach for the cooperative optimization problem would require the development of an analytical wind farm power function which is derived from physical principles, empirical observations and experimental calibrations [42]. However, deriving an analytical wind farm power function is a difficult and complex task that involves uncertainties and assumptions. A data driven approach that bases entirely on the measurement data is one alternative to the power maximization problem.

An experimental study was conducted to test the BA algorithm for the cooperative control optimization problem. As depicted in Fig. 7, wind tunnel experiments with scaled wind turbines were performed at KOCED's Wind Tunnel located at Chonbuk

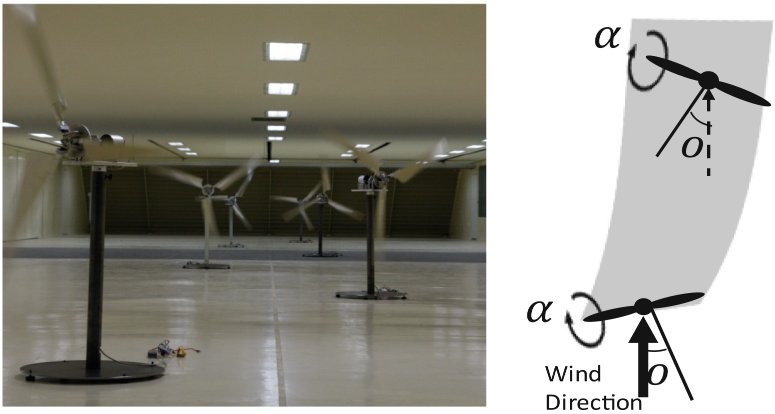
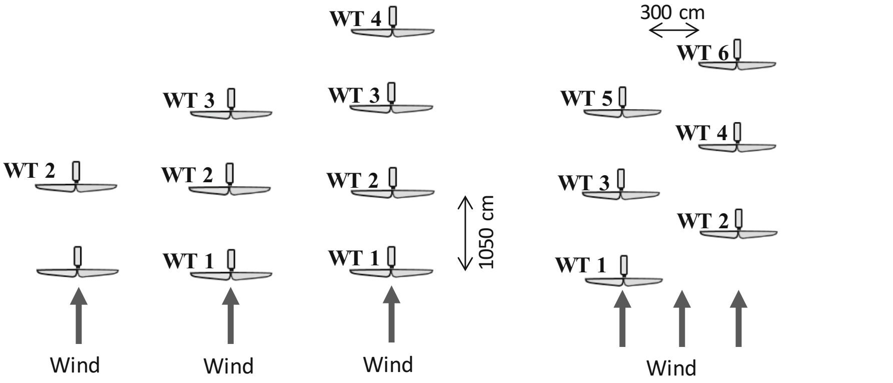


Fig. 7. Experimental wind tunnel set up with scaled wind turbines

National University in Korea [43]. The overall dimension of the wind tunnel is 40 m long, 12 m wide, and 2.2–2.5 m high. The scaled wind turbines are made of three 70 cm long aluminum blades and the rotor diameter is 150 cm. We use the yaw angle O and blade angle α of the wind turbines as the input control actions $\boldsymbol{x}$ and the total power output collected from the wind turbines as the output measurement y. The blades are controlled by a servomotor (Dynamixel-64T) through a mechanical linkage that allows the blade pitch angles α to vary within a 20° range from its original position. The yaw is controlled by the same type of servomotor through a mechanical gear system that allows the yaw angle O rotates from −40° to 40°. The ranges of the servomotors correspond to the physical constraints $\boldsymbol{A}$ imposed on the control inputs $\boldsymbol{x}^l \leq \boldsymbol{x} \leq \boldsymbol{x}^u$. An AC generator is used to convert the mechanical energy into electrical energy.

Multiple experiments with different arrangements of the wind turbines have been conducted. As illustrated in Fig. 8, the longitudinal separation is set at 1050 cm (following the conventional minimum distance of 7 times the rotor diameter between the wind turbines). A constant wind speed is set at 4 m/s (measured at a distance of 32 m from the front of the test section) for all the experiments.

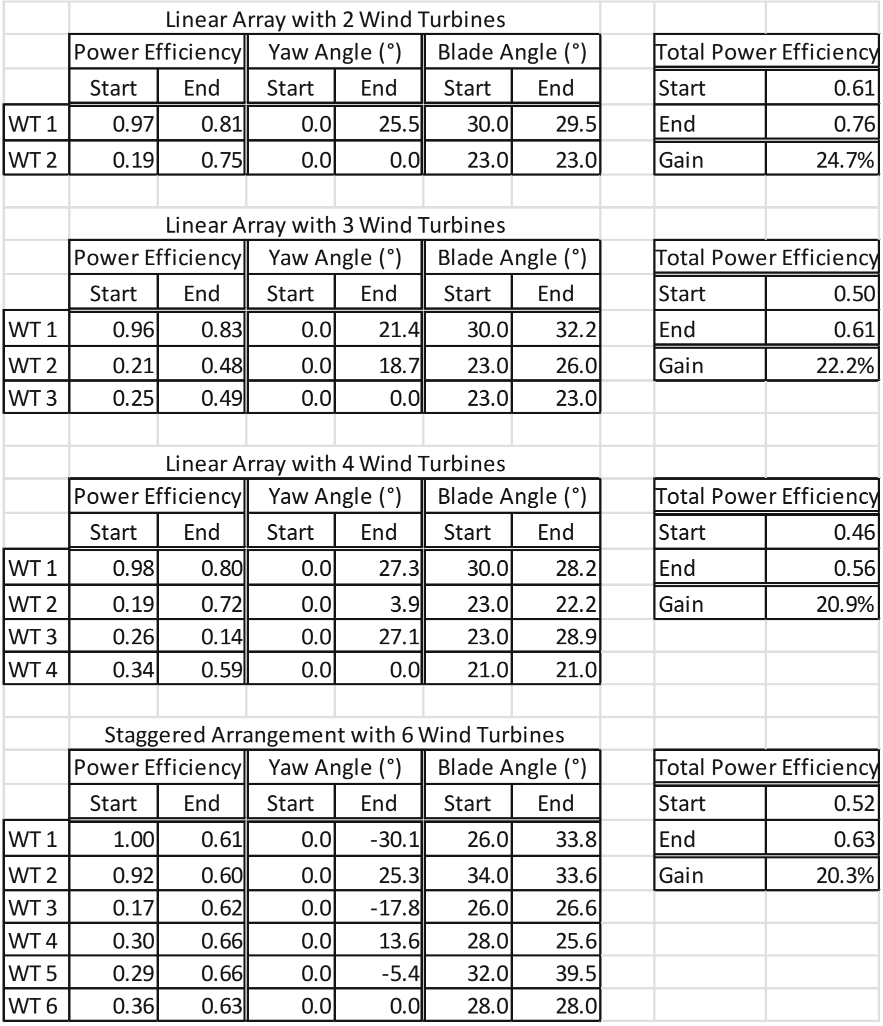


(a) Linear arrays of 2, 3, 4 wind turbines (b) Staggered arrangement of 6 wind turbines

Fig. 8. Four case studies: linear and staggered arrays of wind turbines

Here we show the results for 4 case studies with different arrangements of the wind turbines as shown in Fig. 8. For each case study, three sets of experiments are conducted to assess the performance of the cooperative control of wind turbines using the BA algorithm. First, we measure the maximum freestream power P_i^F of a wind turbine i that can be produced at its location when it operates alone and without wake interference from other wind turbines. From the measurements, the absolute maximum power for all the wind turbines employed in the case study can be computed as $P^F = \sum_i P_i^F$. Second, for comparison, we measure the maximum power P_i^G of a wind turbine i that can be produced at its location when the upstream wind turbines are producing their maximum powers. For this greedy strategy, the wind farm power efficiency relative to the absolute maximum power from the freestream measurements can be computed as $\sum_i P_i^G / P^F$. Finally, the third experiment is conducted to measure the maximum power P_i^C of wind turbine i that is produced at its location when executing the BA algorithm for the cooperative strategy. Starting from the configuration obtained from the greedy control, the BA algorithm is applied to coordinate the actions of the wind turbines as an attempt to maximize the total wind farm power. The wind farm power efficiency for the cooperative control strategy is then computed as $\sum_i P_i^C / P^F$.

During the test, the blade pitch angle and the yaw offset angle of the last wind turbine located at the end of the wind tunnel is fixed in its position. The results measured after 30 iterations for the different case studies are summarized as shown in



Linear Array with 2 Wind Turbines

	Power Efficiency		Yaw Angle (°)		Blade Angle (°)	
	Start	End	Start	End	Start	End
WT 1	0.97	0.81	0.0	25.5	30.0	29.5
WT 2	0.19	0.75	0.0	0.0	23.0	23.0

Total Power Efficiency	
Start	0.61
End	0.76
Gain	24.7%

Linear Array with 3 Wind Turbines

	Power Efficiency		Yaw Angle (°)		Blade Angle (°)	
	Start	End	Start	End	Start	End
WT 1	0.96	0.83	0.0	21.4	30.0	32.2
WT 2	0.21	0.48	0.0	18.7	23.0	26.0
WT 3	0.25	0.49	0.0	0.0	23.0	23.0

Total Power Efficiency	
Start	0.50
End	0.61
Gain	22.2%

Linear Array with 4 Wind Turbines

	Power Efficiency		Yaw Angle (°)		Blade Angle (°)	
	Start	End	Start	End	Start	End
WT 1	0.98	0.80	0.0	27.3	30.0	28.2
WT 2	0.19	0.72	0.0	3.9	23.0	22.2
WT 3	0.26	0.14	0.0	27.1	23.0	28.9
WT 4	0.34	0.59	0.0	0.0	21.0	21.0

Total Power Efficiency	
Start	0.46
End	0.56
Gain	20.9%

Staggered Arrangement with 6 Wind Turbines

	Power Efficiency		Yaw Angle (°)		Blade Angle (°)	
	Start	End	Start	End	Start	End
WT 1	1.00	0.61	0.0	-30.1	26.0	33.8
WT 2	0.92	0.60	0.0	25.3	34.0	33.6
WT 3	0.17	0.62	0.0	-17.8	26.0	26.6
WT 4	0.30	0.66	0.0	13.6	28.0	25.6
WT 5	0.29	0.66	0.0	-5.4	32.0	39.5
WT 6	0.36	0.63	0.0	0.0	28.0	28.0

Total Power Efficiency	
Start	0.52
End	0.63
Gain	20.3%

Fig. 9. Experimental results from executing BA algorithm for cooperative control

Fig. 9. For all cases, significant gain in the total power production can be observed and the maximum gain is obtained. It can be seen that the power productions for the upstream wind turbines are initially near the maximum free stream power efficiencies being more than 95%. Initially, at the start of the cooperative control, the yaws are all directly facing the attacking wind direction (i.e. O = 0) as the result from the greedy control. During the execution of the BA algorithm, the upstream wind turbines offset their yaw angles and reduce their power efficiencies so that the power productions of the downstream wind turbines increase. As a result, the total power production of the wind farm array increases. It is also interesting to observe that, for the staggered configuration with 6 wind turbines, WT 1 and WT 3 offset their yaw angles in clockwise direction while WT 2 and WT 4 offset in counter-clockwise direction such that the wakes are diverted away from the downstream wind turbines.

4 Summary and Discussion

Data-driven machine learning has been an active area of research and has been successfully applied in business, medical, engineering and many other domains. As engineering systems are inherently complex, dynamic, and uncertain, and the data observed in a physical environment is often noisy, probabilistic-based Bayesian learning can play an important role in developing machine learning models. The uncertainties estimated with the learned models can provide valuable insights about the system or problem of interest and help decision making. This paper reviews two Bayesian-based approaches, namely Gaussian Process Regression and Bayesian Optimization, and demonstrates their potential applications for system characterization, diagnostic analysis and control optimization and their implementation with physical engineering systems.

The use of a non-parametric regression model, namely Gaussian Process Regression (GPR), allows the complex relationship between the input features and the target value of the system of interest to be modelled. In this paper, we discuss the use of GPR to develop predictive energy consumption model and predictive tool condition diagnostic model of a milling machine tool. In constructing predictive models useful for practical implementation, it is important to select meaningful features appropriate for the problem of interest.

- To establish the predictive power consumption model for the milling machine, we combine the information collected from a milling machine controller and the power meter. GP is then employed to model the relationship between the input machining parameters and output energy consumption and constructs a prediction function for the energy consumption with confidence bounds. It is important to note that the raw time series data are first transformed into meaningful features that are related to the numerical control (NC) code for the milling operations [32]. The GPR models can thus be used to analyze the milling operations as described by the NC code and to select the best tool path for machining a part [13, 14].
- When establishing the predictive tool condition diagnostic model, sensor (vibration and sound) data that are relevant to understand the tool conditions are collected. The

time series data are pre-processed using the Welch's method that results in smoother power spectrum. Furthermore, the transformation from time series of an arbitrary length data to the frequency domain produces the data sets with specific number of points and reduces dimensionality significantly. The GP model provides confidence bounds for the predictive estimations which can be useful in a practical application where the tool-condition predictions are used to determine when to change machine tools.

For both illustrative case examples, the GPR models are trained with relatively small number of experimental training data sets but the learned model is able to give reasonable predictive estimation quantified with uncertainty measures. As more data become available, the GPR models and their prediction capabilities should improve. On the other hand, one of the drawbacks of the GP, in the form that is described in this paper, is that it uses the whole set of samples or features information to perform the prediction. When the size of training data is large, approximated methods for training and prediction to reduce the computational and storage requirements may be necessary [44–47].

For control optimization, we discuss the Bayesian Ascent (BA) algorithm that optimizes a target system using limited amount of data [25]. The BA algorithm builds upon the Bayesian optimization framework but augmented with a trust region constraint. During the learning phase, the input and output relationship of a target system is modelled using Gaussian Process (GP) regression. At each iterative step, the algorithm exploits the constructed GP model function and determines the next sampling point that can best increase the expected improvement. The trust region constraint ensures that the next input is selected from the region near the best input observed so far. Furthermore, the BA algorithm is able to increase a target value incrementally with gradual changes in the input actions. To illustrate applicability to physical problems, the BA algorithm is employed to the wind farm power maximization problem using only input (control actions) and output (wind farm power production) data. The experimental results show that the BA algorithm is able to increase the total power production by gradually changing the control actions of the wind turbines. Without explicitly constructing the objective function, the BA algorithm is able to optimize the operations of a complex physical system using only the measurement data from the system.

While this paper focuses the discussion on the Gaussian Process Regression and Bayesian Optimization, there exist a broad range of machine learning techniques. It should be noted that selection of an appropriate machine learning technique is problem and data dependent and would require judicious engineering knowledge of the problem involved. As computing and sensing technologies advance, the scope of machine learning tasks will continue to grow. There is no doubt that we will continue to see the impacts of data-driven machine learning in engineering for years to come.

Acknowledgments and Disclaimer. The authors would like to acknowledge the assistance from the late Prof. David Dornfeld of the Laboratory for Manufacturing and Sustainability at UC Berkeley and Dr. Raunak Bhinge of Infinite Uptime, Inc., who conducted the experiments and collected the machine operation data on the Mori Seiki Milling Machine. The authors would also like to thank Prof. Soon-Duck Kwon of Chonbuk National University in Korea, who generously provided the wind tunnel laboratory (KOCED Wind Tunnel Center) facilities for the wind farm power production experiments.

The work described in this paper was partially supported by the National Institute of Standards and Technology (NIST) cooperative agreement with Stanford University (Grant No. 70NANB 12H273 and 70NANB17H031), and the US National Science Foundation (NSF) (Grant No. ECCS-1446330). Any opinions, findings, conclusions or recommendations expressed in the paper are solely those of the authors and do not necessary reflect the views of NSF, NIST and the authors' collaborators. Certain commercial systems are identified in this paper; such identification does not imply recommendation or endorsement by NSF, NIST, Stanford University or the authors; nor does it imply that the products identified are necessarily the best available for the purpose.

References

1. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives and prospects. Science **349**, 255–269 (2015)
2. Wuest, T., Weimer, D., Irgens, C., Thoben, K.-D.: Machine learning in manufacturing: advantages, challenges, and applications. Prod. Manuf. Res. An Open Access J. **4**, 23–45 (2016)
3. Suresh, P.V.S., Venkateswara Rao, P., Deshmukh, S.G.: A genetic algorithmic approach for optimization of surface roughness prediction model. Int. J. Mach. Tools Manuf. **42**, 675–680 (2002)
4. Ghosh, N., Ravi, Y.B., Patra, A., Mukhopadhyay, S., Paul, S., Mohanty, A.R., Chattopadhyay, A.B.: Estimation of tool wear during CNC milling using neural network-based sensor fusion. Mech. Syst. Sign. Process. **21**, 466–479 (2007)
5. Ak, R., Helu, M., Rachuri, S.: Ensemble neural network model for predicting the energy consumption of a milling machine. In: 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2015), p. V004T05A056. ASME (2015)
6. Jihong, Y., Lee, J.: Degradation assessment and fault modes classification using logistic regression. J. Manuf. Sci. Eng. **127**, 912–914 (2005)
7. Carbonneau, R., Laframboise, K., Vahidov, R.: Application of machine learning techniques for supply chain demand forecasting. Eur. J. Oper. Res. **184**, 1140–1154 (2008)
8. Ghahramani, Z.: Probabilistic machine learning and artificial intelligence. Nature **521**, 452–459 (2015)
9. Orbanz, P., Teh, Y.W.: Bayesian nonparametric models. In: Encyclopedia of Machine Learning, pp. 81–89. Springer, New York (2011)
10. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
11. Rasmussen, C.E.: Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) ML-2003. LNCS (LNAI), vol. 3176, pp. 63–71. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28650-9_4
12. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
13. Bhinge, R., Park, J., Law, K.H., Dornfeld, D., Moneer, M., Rachuri, S.: Towards a generalized energy prediction model for machine tools. J. Manuf. Sci. Eng. **139**(4), 041013 (2017)

14. Park, J., Law, K.H., Bhinge, R., Biswas, N., Srinivasan, A., Dornfeld, D., Helu, M., Rachuri, S.: A generalized data-driven energy prediction model with uncertainty for a milling machine tool using Gaussian Process. In: 2015 International Manufacturing Science and Engineering Conference (MSEC 2015), p. V002T05A010. ASME (2015)
15. Nguyen-Tuong, D., Peters, J.R., Seeger, M.: Local Gaussian process regression for real time online model learning. In: 22nd Annual Conference on Neural Information Processing Systems. Advances in Neural Information Processing Systems 21, pp. 1193–1200 (2008)
16. Teramura, K., Hideharu, O., Yuusaku, T., Shimpei, M., Shinichi, M.: Gaussian process regression for rendering music performance. In: International Conference on Music Perception and Cognition (ICMPC 10), pp. 167–172 (2008)
17. Alpaydm, E.: Introduction to Machine Learning, 3rd edn. MIT Press, Cambridge (2014)
18. Brochu, E., Cora, M.V., Freitas, N.: A tutorial on Bayesian optimization of expensive cost functions with application to active user modelling and hierarchical reinforcement learning. Technical report, University of British Columbia, Canada (2010). arXiv:1012.2599
19. Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. J. Global Optim. **13**, 455–492 (1998)
20. Osborne, M.: Bayesian Gaussian process for sequential prediction, optimization and quadrature. Ph.D. Dissertation, Department of Computer Science, University of Oxford, UK (2010)
21. Bubeck, S., Munos, R., Stoltz, G., Szepesvari. C.: X-armed Bandits. J. Mach. Learn. Res. **12**, 1655–1695 (2011)
22. Scott, S.L.: A modern Bayesian look at the multi-armed bandits. Appl. Stochast. Models Bus. Indus. **26**(6), 639–658 (2010)
23. Osborne, M., Garnett, R., Roberts, S.: Active data selection for sensor networks with faults and changepoints. In: IEEE International Conference for Advanced Information Networking and Applications (2010). https://doi.org/10.1109/aina.2010.36
24. Garnett, R., Osborne, M., Roberts, S.: Bayesian optimization for sensor set selection. In: Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Network, pp. 209–219 (2010)
25. Park, J., Law, K.H.: Bayesian ascent: a data-driven optimization scheme for real-time control with application to wind farm power maximization. IEEE Trans. Control Syst. Technol. **24** (5), 1655–1668 (2016)
26. Duvenand, D.K.: Automatic model construction with Gaussian processes. Ph.D. Thesis, University of Cambridge (2014)
27. Genton, M.G.: Classes of kernels for machine learning: a statistics perspectives. J. Mach. Learn. Res. **2**, 299–312 (2001)
28. Rasmussen, C.E., Nickisch, H.: Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. **11**, 3011–3015 (2010)
29. Blondel, M., et al.: Scikit-learn, machine learning in Python. http://scikit-learn.org/stable/. Accessed 22 Dec 2016
30. Sobel, W.: MTConnect Standard. Part 1—Overview and Protocol, Version 1.3.0 (2015)
31. Vijayaraghavan, A., Dornfeld, D.: Automated energy monitoring of machine tools. CIRP Ann. Manuf. Technol. **59**, 21–24 (2010)
32. Bhinge, R., Biswas, N., Dornfeld, D., Park, J., Law, K.H., Helu, M., Rachuri, S.: An intelligent machine monitoring system for energy prediction using a Gaussian process regression. In: IEEE International Conference on Big Data, pp. 978–986. IEEE (2014)
33. Fan, J., Han, F., Liu, H.: Challenges of big data analysis. Natl. Sci. Rev. **1**(2), 293–314 (2014)
34. Kannatey-Asibu, E., Dornfeld, D.A.: A study of tool wear using statistical analysis of metal-cutting acoustic emission. Wear **76**(2), 247–261 (1982)

35. Dimla, D., Lister, M.: Online metal cutting tool condition monitoring - I: force and vibration analyses. Int. J. Mach. Tools Manuf. **40**(5), 739–768 (2000)
36. Stoica, P., Moses, R.L.: Spectral Analysis of Signals, vol. 452. Pearson Prentice Hall, Upper Saddle River (2005)
37. Allen, R.L., Mills, D.: Signal Analysis: Time, Frequency, Scale, and Structure. Wiley, New York (2004)
38. Welch, P.: The use of Fast Fourier Transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967)
39. Williams, C.K., Rasmussen, C.E.: Gaussian processes for regression. In: Advances in Neural Information Processing Systems, pp. 514–520 (1996)
40. Wilson, A.G., Adams, R.P.: Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, vol. 28, pp. III-1067–III-1075 (2013)
41. Mockus, J., Fretitas, A., Castelanous, J.A.: Toward Global Optimization. North-Holland, Amsterdam (1978)
42. Park, J., Law, K.H.: Cooperative wind turbine control for maximizing wind farm power using sequential convex programming. Energy Convers. Manag. **101**, 295–316 (2015)
43. Park, J., Kwon, S., Law, K.H.: A data-driven, cooperative approach for wind farm control: a wind tunnel experimentation. Energies **10**, 852 (2017)
44. Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. In: Neural Information Processing Systems (NIPS) Conference, pp. 1257–1264 (2005)
45. Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. **6**, 1939–1959 (2005)
46. Ranganathan, A., Yang, M.H., Ho, J.: Online sparse Gaussian process regression and its applications. IEEE Trans. Image Process. **20**(2), 391–404 (2011)
47. Park, J., Bhinge, R., Law, K.H., Dornfeld, D., Mason, C., Rachuri, S.: Real time energy prediction for a milling machine tool using sparse Gaussian process regression. In: International Conference on Big Data, pp. 1451–1460 (2015)

3D Imaging in Construction and Infrastructure Management: Technological Assessment and Future Research Directions

Yujie Wei[1], Varun Kasireddy[1], and Burcu Akinci[2(✉)]

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA
yujiew@andrew.cmu.edu, varunkasi@cmu.edu
[2] Paul P. Christiano Professor, Carnegie Mellon University, Pittsburgh, PA 15213, USA
bakinci@cmu.edu

Abstract. With rapid developments in 3D imaging technology, as well as the evolving need in Architecture/Engineering/Construction/Facility Management (AEC/FM) industry to understand various field aspects from a 3D perspective, several technologies, such as laser scanning, camera and RGBD camera, are becoming important components of civil engineers' and architects' toolbox. With improvements in efficiency and reduction in operational cost, new avenues are opening in leveraging 3D imaging sensors for construction and infrastructure management. In the light of this, there is a need to assess the achievements to date, especially with respect to unique challenging use case scenarios and requirements that construction and infrastructure management domains provide, and consequently identify potential research challenges that still need to be tackled. This paper targets doing such an assessment around several challenges faced when using 3D imaging to support construction and infrastructure management. It specifically discusses to what extent these challenges have been addressed and several approaches that are used in addressing them. It also presents current 3D imaging development trends and discusses briefly some challenges that are emerging due to increased application of 3D imaging techniques to highlight future research needs in this domain.

Keywords: Construction and infrastructure management · 3D imaging
Point cloud · Laser scan · Building information modeling · Automation
Sensors

1 Introduction

Over the past decade or so, the nature of construction and infrastructure management is shifting towards a paradigm in which there is an increasing demand for understanding various aspects of what is happening at the field from a 3D perspective. Hence, 3D imaging sensors, such as LiDAR, camera, and RGBD camera, are becoming parts of the common toolbox of civil engineers and architects in many Architecture/Engineering/Construction/Facility Management (AEC/FM) research projects and real-world case studies [e.g., 1–3]. Several example applications as a result of leveraging such

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 37–60, 2018.
https://doi.org/10.1007/978-3-319-91635-4_3

technologies range from, 3D reconstruction for as-is modeling [4–10], progress monitoring [11–13], safety management [14–16] and inspection and quality control [17–19]. In the past 20 years, as 3D imaging technologies continued to evolve, several workflows have been established for these applications to satisfy various requirements in different usage scenarios. This has consistently resulted in improved efficiency for sensing and modeling a scene, and reduction in operational cost of documenting and maintaining a built environment. Given recent rapid improvements in sensing technologies, several researchers are embracing the usage of such technologies in new applications, such as accessibility diagnosis [20] and 3D thermal mapping [21], and they are also increasing the scalability of the existing applications [22]. This results in new avenues opening up in leveraging 3D imaging sensors for construction and infrastructure management. Given all these developments, there is a need for assessing achievements to date, especially with respect to unique challenging use-case scenarios and requirements that construction and infrastructure management domains have. In addition, it is also important to identify potential research challenges that still need to be tackled. Therefore, this paper assesses achievements to date and identifies future research agenda for 3D imaging to support construction and infrastructure management.

Several previous research studies have reviewed existing 3D imaging technologies within the AEC/FM context from two perspectives: *workflow* and *application*. The reviews from the *workflow* perspective typically proposed overarching end-to-end processes across different workflows from 3D imaging to decision support and discussed state-of-the-art practices within those processes [23–25]. The discussion on workflows has gradually converged to a consistent point cloud-centric one [23, 26–28], which divides the 3D imaging workflow into three phases: point cloud or imagery data capturing, point cloud generation, preprocessing, and modeling. On the other hand, previous research studies focusing on reviewing literature from an *application* perspective usually targeted a specific usage scenario, such as as-built modeling [18, 26, 28, 29], safety management [30], construction monitoring [25, 31, 32] and defect detection [33]. These reviews first outlined currently practiced workflows done in a traditional manner (i.e. without using 3D imaging technologies), and then proposed a 3D imaging-based workflow with respect to a specific usage scenario. However, in all these studies, there lacks a synthesis of 3D imaging technologies driven by some specific challenges faced within the AEC/FM community targeting the following questions: What are primary challenges of applying 3D imaging techniques that the AEC/FM community has been facing in the past 20 years? What achievements have been accomplished to-date and how have these primary challenges been addressed? Moreover, what are some emerging challenges as 3D imaging technologies and their usages have evolved over time?

This paper first synthesizes widely accepted challenges presented in current practice across different AEC/FM applications, such as occlusions, level of detail, lack of 3D features, feature-less structure, distance and speed of data capture, data accuracy, and interoperability and maintainability of a model. Some of these challenges have been well-addressed in prior research studies. In this regard, the second and the largest part of the paper presents a synthesis on how these challenges have been addressed in state-of-the-art 3D imaging practices. Specifically, the second part of the paper aims at answering the question of how state-of-the-art practices address these challenges respectively. Following the discussion of achievements, the third part of the paper focuses on discussing future trends and emerging new challenges. The categories of trends that are observed: (1) from static to mobile scanning, (2) from deterministic to probabilistic methodology for data processing, and (3) from using a single type of sensor to multi-sensor fusion to increase accuracy and coverage. The authors also identified some emerging challenges with respect to each trend and outlines open research questions with regards to each challenge.

2 Challenges Towards Robust and Semantic 3D Imaging

This section discusses some of the primary challenges existing in the current practice of using 3D imaging technologies within the AEC/FM industry.

Occlusions and Clutter. Occlusions block a visual sensor from seeing the object of interest, requiring either a sensor to bypass them or utilization of an algorithm to infer a complete object behind occlusions. Many AEC/FM applications are susceptible to occlusions, due to several reasons, such as: (1) dense deployment of infrastructure components in a given scene, e.g., bracings under a bridge (at least 35–40% of the deck underside is occluded in Fig. 1) and pipelines crammed in an equipment room, (2) temporary components during construction, such as scaffolding and equipment, (3) mixture of moving components (such as appliances and furniture) and fixed architectural/infrastructure components.

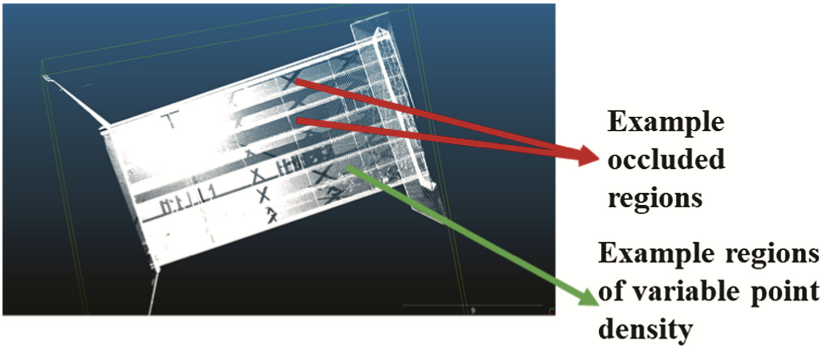


Fig. 1. Example image from under a bridge showing missing data due to occlusion (x-bracings and girders) as well as the varying density of points in certain regions compared to their neighboring regions

Another issue with point cloud space is clutter, which results in needing to adjust the level of detail of data capture and modeling as per specific use-case scenarios. For example, it is possible to see and capture each brick on a brick wall in 3D capture technologies, depending on the need for modeling details, the output model can be as simple as a wall object or a masonry wall consists of bricks. Similarly, there are temporary objects in a scene (e.g., formwork, scaffolding, furniture, etc.), that 3D imaging technologies capture in a scene and might not be relevant for many uses case, such as construction quality control. Hence, it is necessary to identify objects of interest and remove objects of no interest, and this is one of the challenges that AEC/FM project conditions pose to the utilization of 3D imaging technologies.

Lack of Features for Understanding 3D Scenes. The goal of scene understanding is to take 3D imaging data, such as RGB(D) image or point cloud, as input, and output semantic understanding of a scene including: objects present in a scene, attributes of those objects, and relationships amongst them. Scene understanding analysis can be divided into two levels: (1) low-level semantic analysis, which focuses on segmenting and clustering imaging data based on similarity, such as grouping neighboring points with similar color into a patch or segmenting a plane from point cloud using normal-based region growing [34], and (2) high-level semantic analysis, which aims at understanding what infrastructure components exist in a scene and how they are related to each other, such as recognizing mechanical-electrical-plumbing (MEP) components.

In this paper, any information, which can be used in a computer-interpretable manner to support *scene understanding* analysis, is referred to as a feature. Features can be broadly divided into two categories based on the way they are computed. Handcrafted features refer to the features extracted through filters identified by a person, such as edges, corners, and SIFT descriptors [35]. The process of identification of a filter incorporates an engineer's prior knowledge and understanding of a scene. Non-handcrafted features are extracted through filters learned from labeled data directly. Figure 2 shows a comparison between handcrafted features and non-handcrafted features. For 2D images, both handcrafted features and non-handcrafted features have been widely used in image classification, object detection, and scene understanding. However, semantic understanding using 3D imaging still heavily depends on handcrafted features only. Identification of feature filters from labeled data using machine learning for 3D imaging representation is still an active research area since: (1) 3D imaging data, especially point cloud, is usually unordered, sparse, and non-uniformly distributed in space, which causes difficulty for convolving a point cloud; (2) 3D imaging data has an extra dimension which requires significantly larger memory and more computation when extracting features ($O(N^3)$ compared to $O(N^2)$ in the 2D space where N is the number of pixels or points); (3) Architectural scenes usually have many texture-poor surfaces (walls, ceilings, floors) and repetitive components (columns, beams) that makes the extracted features non-unique and complicated for visibility reasoning [36]. The quality of the reconstructed point cloud also gets affected by having lack of features in a scene, as shown in Fig. 3. The so-called perceptual aliasing problem, the phenomenon in which two or more different scenes can be represented with the same 3D imaging data, makes it difficult for learning features purely from data. (4) Feature learning is data-hungry. For instance, the ImageNet dataset contains over 10 million labeled images. While several researchers have started sharing their datasets to help improve 3D scene understanding in the AEC/FM domain [37, 38], in most of the domain-specific cases, researchers still work on their own datasets, which may not be accessible to other researchers. This makes it difficult to compare the performance of algorithms developed by different researchers, as well as to assess which research directions can effectively accelerate 3D scene understanding. For this reason, it is important to regularly direct efforts towards creating 3D benchmark datasets and organizing hackathons, much like the efforts observed the traditional Computer Vision and Machine Learning communities, such as ImageNet [39] and COCO [40].

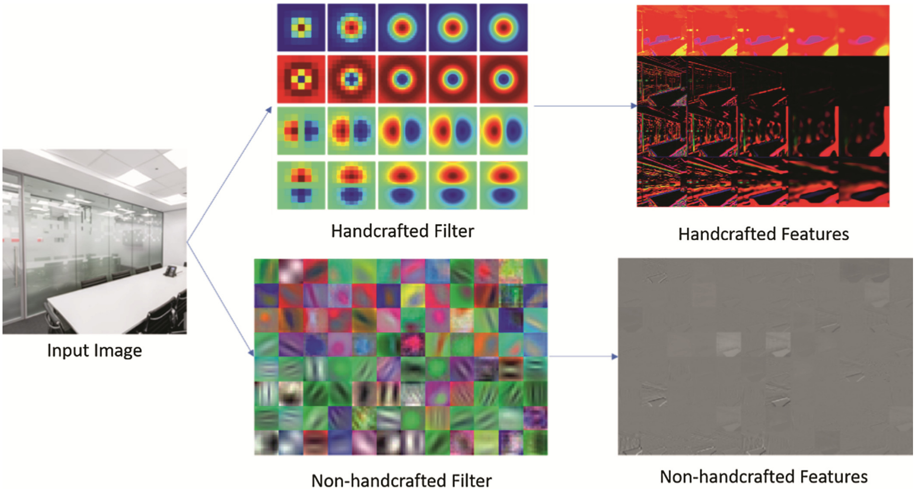


Fig. 2. Comparison between handcrafted features and non-handcrafted features. The handcrafted filters [41] above are designed manually while the non-handcrafted filters [42] below are learned from data.

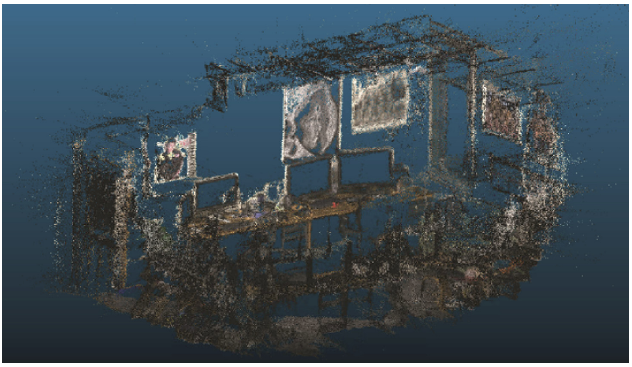


Fig. 3. Reconstructed point cloud from images missing the surface components due to poor texture and features on walls.

Distance, Speed and Accuracy. Requirements associated with needing to capture data from longer distances and faster speed arise from increasing scale and complexity of a scene [36]. However, this typically comes with a tradeoff of reduction in accuracy. The tradeoff between accuracy and speed is becoming even more critical when considering robustness, especially with respect to wide a variety of scan environments. For example, although modern LiDAR can capture a relatively dense point cloud from over 100 m away with an accuracy of up to 2 mm or less [43], accuracy drastically reduces when scanning glass structure or other reflective surfaces in bright conditions [44]. Sometimes, a workaround in such a situation is to scan the same area multiple times and take the average of the measurements. However, such operation can drastically slow down data capture process, and can be untenable in large scenes. In contrast, while photogrammetry is convenient in most cases, it is still impacted by environmental conditions, such as motion blur, lighting condition, and camera calibration [45].

Interoperability. Interoperability, in the context of the 3D imaging industry, means enabling different components of a technology (both hardware and software) to work together. Sometimes, it involves communication between: (i) two or more hardware systems (scanners and on-site laptops that are used for tuning scan settings and preprocessing), (ii) hardware and software (scanners from different scanner vendors and point cloud processing software, such as Autodesk ReCap), and (iii) two or more software (preprocessing software, such as Autodesk ReCap, and modeling and analysis software, such as Autodesk Revit), in order to facilitate and streamline various tasks, such as data collection, modeling and analysis. More specifically, interoperability measures the difficulty/ease of: (i) different hardware systems interacting through a communication protocol, and (ii) transforming data from one format to another format based on application requirements. Lack of communication will hinder the ease of use and operation of a specific 3D imaging technology. Whereas, lack of interoperability of data can reduce users' ability to fully exploit all possible information contained in a collected data set, which in turn can impede the effectiveness and efficiency of decision making [46].

3 Achievements in Addressing Existing Challenges

This section provides a synthesis on how recent developments in 3D imaging techniques aim at addressing widely accepted challenges presented across different applications. Specifically, this section discusses existing state-of-the-art methods to address the specific challenges highlighted in Sect. 2. Table 1 gives a summary of the achievements-to-date and the subsections below describe the existing approaches and achievements in more detail.

Table 1. 3D imaging achievements with regard to challenges stated in Sect. 2

Challenges	Proposed approaches	Achievements within each approach and across different approaches
Occlusion and clutter	Path/time/scan planning	Path planning considering visibility coverage [47, 48], budget [49], or safety [47]
	Mobile systems	Mobile laser scanning [56], integrated systems, such as backpack and drones [57]
	Detection and removal of unnecessary objects in a scene	Knowledge-based occlusion detection and removal [52–54].
	Human-computer interaction approaches to enable users to make decisions	Occlusion-aware interface design that lets the user handle the occlusion [55]
3D features	Domain-specific hand-crafted features	Utilization of hand-crafted features for civil-related object recognition, such as scaffolding [4] and MEP components [58]
	Sensor fusion	Integrated systems for automatic infrastructure mapping [59], inspection [57, 60], and construction monitoring [32]
	2D Conversion	Applied image processing techniques on voxelized [61] or projected point clouds [62] to understand a scene
	General 3D features	Directly used deep neural network to extract general 3D features using symmetric functions [63]
Trade-off between accuracy, distance, and speed	Formalized approaches that would enable choosing of different systems based on application requirements	Emergence of image-based methods for modeling [36, 64], inspection [19], monitoring [65], and localization [66, 67]
		Incorporation of accuracy, quality, and speed trade-off with respect to specific application [68–70]
		Automatic preprocessing of the point cloud to improve the speed and quality [71]
Interoperability	Data format interoperability	Standard format of 3D imaging data [72, 73]
	Interoperability for modeling tasks	Standard-shape modeling based on IFC [74]

3.1 Occlusions

The current practice of handling occlusions can be generally divided into two categories: (1) Occlusion elimination/minimization, and (2) Enabling occlusion awareness.

1. *Occlusion elimination/minimization methods* further consists of the following:

Path/Time Planning Approaches. These approaches target designing/planning a scanning route/schedule to get better coverage of a scene. For example, one typical goal of path/schedule planning approach within this context is to find an optimal path/schedule that maximizes the visibility of objects of interest, with [47] or without [48] scene information. Other criteria, such as budget [49] and safety [47], were also considered in existing path planning approaches. In an active construction quality control system [50], a routine inspection plan was proposed to provide consistent monitoring of structural components while mitigating the effect of random occlusions in a single inspection.

Utilization of Mobile Sensors to Support Flexible and Adaptable Data Capturing. The mobile sensor can either be as simple as a handheld laser scanner with a calibrated camera or as complicated as an integrated system like a backpack [51] or a robot [47]. These integrated systems can be operated by a human being or can explore a scene autonomously. Real-time visualization of the surrounding environment allows an operator to spot occlusions and bypass them easily. Active detection and elimination require high-level semantic understanding because related algorithms have to distinguish unnecessary objects in a scene [52–54].

2. *Occlusion awareness methods* aim at enabling users to be aware of existing occlusions and deal with them through better interface design and probabilistic methods [55]. This will help in determining if the data currently at hand is sufficient to make a specific decision, or if more data needs to be collected.

3.2 Lack of Effective 3D Features for Scene Understanding

Data analysis typically involves two primary tasks: Registration and Feature Extraction. Registration aims at estimating a transformation (translation and rotation) that puts different sets of data into a single coordinate system. For instance, Iterative Closest Point (ICP) registration iteratively minimizes the Euclidean distance between two point clouds to align them in the same coordinate system. Feature extraction turns raw 3D imaging data (pixel and point intensity) into higher-level semantic information (contrast, color difference, curvature, etc.). The extracted feature from 3D imaging data can be geometric elements, such as edges, corners, and planes. Other examples of extracted features are related to some properties of a region, such as color distribution within a locally textured patch in a scene.

As discussed in Sect. 2, point cloud processing still heavily relies on handcrafted features that require prior knowledge of a scene. The prior knowledge, on one hand, is very helpful for a specific problem. For instance, when detecting floors from a point cloud, it is reasonable to assume that all floors are planes that are perpendicular to the gravity direction. Using normal direction and difference of normal (DoF) within a local region as a feature, only the points with a vertical normal direction and a small DoF will be labeled as a floor point. On the other hand, strong prior knowledge limits the generalization ability of a handcrafted feature, i.e., the performance of a feature in an unexpected scenario. In the floor detection example above, the normal direction feature will fail when encountering a slope. To support semantic understanding of a scene, a feature must present strong ability of generalization. One metric of measuring the generalization ability of a feature is the number of objects it can distinguish. For example, using normal direction only, a program can identify the difference between walls and floors, but cannot distinguish a door and a wall because they have similar normal directions. Using the number of distinguishable objects as a metric, we observed that effective features for 3D imaging is limited in identifying many different types of objects existing in an AEC/FM scene. For example, a typical point cloud classification dataset Sementic3D [75] contains 8 classes of objects (man-made terrain, natural terrain, etc.), while the most famous image dataset, ImageNet [39], contains over 20,000 categories ranging from animals to fungus.

Many approaches have been proposed to address the need for effective 3D features for point-clouds, such as (1) domain-specific features, (2) sensor fusion, (3) feature extraction on 2D images, and (4) 3D learned features. Below are discussions with respect to each approach respectively.

Domain-Specific Features. Previous research focused on developing domain-specific hand-crafted features using local descriptors, such as Fast Point Feature Histogram (F/PFH) [76] or global descriptors, such as Fast Viewpoint Feature Histogram (F/VFH) [77]. The PFH is a local descriptor that captures geometric information of surrounding points and it measures differences amongst the normal of points within a vicinity of a point or a cluster of points. A VFH is a local descriptor of a point cloud that measures the geometric information of a patch when the patch is viewed from a certain angle. For example, a column usually shares a similar width when viewed from different directions, but a wall has a larger width when it is being viewed from the front. Depending on the size of an object, it is difficult to choose a proper vicinity when computing a VFH. In practice, a VFH is often calculated at different resolution scales and concatenated to improve the robustness of an algorithm to local variations The local descriptors mentioned above are useful for identifying some specific components, such as scaffolding [4] and Mechanical/Electrical/Plumbing components [58].

Sensor Fusion to Leverage 2D Features. Sensor fusion registers data captured by multiple sensors into a single coordinate system and enables taking advantage of specific data sets provided by each sensor. Typical sensor fusion approaches include fusion of camera and LiDAR, and camera and LiDAR and localization sensors such as GPS and IMU [59, 60].

Registering 2D images with 3D point clouds provides many benefits. First, 2D-3D fusion is a useful approach to generate 3D models with textures for as-built infrastructure [78]. Second, 2D-3D fusion can mitigate several weaknesses of each sensor. For instance, 3D point interface, although more accurate from a measurement perspective, does not provide an easy-to-use visual interface. On the other hand, 2D images are visually more convenient to handle, but suffer from scaling and perspective issues [79]. Therefore, 2D-3D fusion can better support many applications that require accurate spatial information and easy-to-use visual information, such as defect measurement and annotation [79]. Through image calibration [80] applied to fused point cloud and image data, it is possible to accurately project 3D point cloud to 2D images. Other approaches to conducting 2D-3D fusion include maximizing the mutual information between images and a point cloud [78] or matching edges and planes [81].

Utilization of localization sensors together with a camera and a laser scanner enables capture and registration of imaging data in real-time. This alleviates the need for off-site processing necessary to get a registered 3D point cloud. Having such unified 3D point cloud available in real time can help in assessing the completeness of data, which in turn can help in minimizing occlusions, by enabling users to intervene and modify a data collection process in real time. It is observed that the fusion of localization sensors and perception sensors leads to the development of integrated systems such as handheld scanners (Faro Scanner Freestyle, Kaarta Stencil, etc.), backpacks [59], and drones [57].

2D Conversion. Through voxelization and projection, 3D point cloud can be converted to image-like representations to enable application of existing 2D feature extraction techniques. Voxelization first divides the whole space into many subspaces called voxels. Each voxel will be labeled using binary representation (occupied/empty) or quantitative representation (number of points, density, or other values). After voxelization, point cloud representation becomes ordered and dense, and therefore many 2D processing methods, such as 3D Convolutional Neural Network (CNN) [82], can be directly applied to voxelized point clouds. Another advantage of 2D conversion methods is that it allows transfer learning. Transfer learning is an algorithm that allows the knowledge learned from other scenarios to be applied to a given scenario. For instance, a typical application of transfer learning is to learn the convolutional neural network parameters from a large set of images, such as ImageNet [39] and fine-tune the network with respect to a certain application [83]. Since most deep learning methods are data-hungry, transfer learning can reduce the need for a large amount of data and improve the generalizability of the 3D imaging techniques.

Automatic 3D Feature Extraction. Instead of converting a point cloud to 2D-similar representations, such as spin image and voxel-based representations, another approach to obtain useful scene understanding features is to learn 3D features from a labeled point cloud directly. These methods employ symmetric operations, such as local maximization (select the maximum element from a vector) or calculating Euclidean norms (the magnitude of a vector) when extracting features. Symmetric operators are independent of input order and therefore these methods are able to take a point cloud as a whole without considering point order or sparseness [84]. With the ability to handle raw point clouds, it is possible to train a model that takes a raw point cloud as input and obtain any desired output, such as segmented point clouds and detected objects and their locations.

3.3 Quality and Efficiency of Data Capture

Though 3D imaging researchers and technology providers continue to work on improving quality and efficiency of data capturing, the tradeoff between these two parameters is still a critical issue to consider when designing a workflow for data collection and analyses to support a specific use-case. Table 2 listed available metrics for evaluating the quality of a point cloud.

Table 2. Point cloud quality metrics.

Metric	Definition	Factors that affect the metric
Location Error	The average distance between the scanned point and the ground truth point	Hardware, registration, calibration
Density	The average number of points per cubic meter	Hardware
Coverage	The ratio between the covered region and the region of interest	Scanning path, occlusions, feature richness of a scene
Feature Richness	The variance of extracted features	Color, material, and geometric difference presented in a scene, hardware

Location error is the distance between a scanned point and its corresponding ground truth. Existing static laser scanners, such as Faro S350 and Leica P50, can conduct range measurements up to several hundred meters with a location error less than 1 mm, at the cost of about a hundred thousand US dollars. RGB-D sensors, such as Microsoft Kinect, support measurements up to 5 m with an accuracy of 4 cm. The depth measurement error of RGB-D sensors increases quadratically with increasing measurement distance [70]. *Density* usually refers to the average number of points within a unit space. It is worth mentioning that the density of a point cloud also decreases quadratically as measurement distance increases. Static laser scanners can capture millions of points per scan, while mobile laser scanners, such as Velodyne HDL64E S2, can only capture about eighty thousand points per frame. When compared to images, the density of a point cloud can be viewed as the reciprocal of the gap between two measurements. Image is therefore considered as a denser representation compared to point cloud because no gap exists between two pixels on an image. *Coverage* measures how well the data captures the object of interest, which depends on data capturing events. *Feature richness* measures the amount of information captured by the data, which is determined by hardware and the texture presented in a scene at the same time. While most static laser scanners can capture color point clouds through built-in cameras, mobile laser scanners can only capture point intensity and location without any color information [85, 86]. Improving data quality with respect to one or more metrics mentioned above requires extra time and money cost. Therefore, it is necessary to understand the data quality requirement of an application and choose a proper workflow.

Existing efforts within the 3D imaging community towards improving data quality and efficiency can be categorized into two directions: *Hardware-specif*ic and *Software-specific*. From hardware perspective, stationary laser scanners are usually able to capture more accurate and richer data compared to mobile laser scanners [45]. In comparison to photogrammetry, mobile laser scanners capture more accurate spatial data with less visual data [69, 87]. From software perspective, achievements in improving captured data quality can be categorized into three: noise reduction, accurate registration, and interpolation. Noise reduction aims at removing outliers or making users aware of outliers to avoid propagating error to the subsequent processing steps. Accurate registration can reduce possible errors that can arise due to imperfect point cloud registration. For example, registration approaches using Sphere Feature Constraints can improve the mean error from 0.1 mm level to 0.01 mm level compared to the ordinary ICP [88]. Interpolation methods, such as depth estimation and dense reconstruction, can improve point cloud density by leveraging other visual features, such as matched keypoints or color consistency. For instance, *multi-view stereo* (MVS) reconstruction, which takes location and orientation information - that is extracted when *structure from motion* (SFM) step is applied on 2D images, can be used to make a 3D dense point cloud. Results from previous research [89], show that point cloud coverage can be improved from less than 10% (right after structure-from-motion reconstruction) to about 98% while keeping the averaged accuracy less than 1 mm.

3.4 Interoperability

In Sect. 2, it is mentioned that improvements in interoperability can facilitate and streamline various tasks, such as data collection, modeling and analysis. Several case studies have been performed in the past decade investigating the interoperability in the context of construction, infrastructure and facility management [90].

In terms of data collection efforts, currently, scan data collected from laser scanners of different vendors can be converted to industry-driven standardized formats such as E57 [91], which in turn can be used for downstream tasks, such as modeling and analysis. In the case of images, it is helpful to collect metadata pertaining to each image, such as position, pose, and units, in order to get a better 3D perception from analyzing those images. Currently, some industry standards exist for defining the format in which image metadata should be collected and organized such as ROS bags [72].

Some case studies discussed interoperability issues amongst different modeling software systems [92]. While modeling standard-shaped components, such as prismoidal walls and columns, from point clouds has been largely addressed [55], problems persisted when modeling some infrastructure systems, such as bridges. For example, a bridge typically carries a roadway. The roadways maybe curvy in three planes, thus adding complexity to the modeling task. Moreover, in some cases, such as damage detection and condition assessment, shapes of defects and damages need to be modeled and reasoned about, in addition to modeling infrastructure components. Although some difficulties associated with geometric modeling have been addressed by some software systems [93, 94], progress in generation and sharing of infrastructure-specific libraries of intelligent and parametric 3D object families has not happened at the same pace. Such libraries could facilitate bi-directional translation between these complex geometries and IFC-inspired representations (e.g. IFC-Bridge in the case of bridges). Further research in this area could improve ease and comprehensiveness in modeling and hence improve interoperability.

Various analyses tasks often involve comparisons between a given 3D model with point cloud data by means of superimposing one on top of the other. This is computationally intensive, due to the large size of point cloud data. It was almost impossible to handle large datasets (~0.5–2 GB) a few years back in a desktop environment (<2 GB GPU capacity), and the workaround required splitting both a model and a point cloud into smaller sections and analyzing them separately. This resulted in inefficient and error-prone analysis [17, 95, 96]. However, with improvements in hardware technologies, such as with GPUs and solid state devices (SSD), as well as through cloud technology, wherein software is offered as Software-As-A-Service (SAAS) [97], it is now possible to analyze medium to large-size datasets (~0.5–2 GB) in a reasonable time frame (up to 5–6 h).

4 Discussion on Trends and Emerging Challenges

In this section, the authors present three observed trends of technological development in the past twenty years: (1) from static to mobile scanning, (2) from deterministic to probabilistic methodology for analyses, and (3) from a usage of single type of sensor to multi-sensor fusion. The sections below overview these trends in terms of achievements to date and highlight remaining and emerging challenges.

4.1 From Static to Mobile Data Capture

Recently, data capture technologies have been migrating towards more mobile platforms, wherever accuracy requirements are not as stringent like mm level accuracy, and in situations where scene components are not easily accessible and visible from fix locations (e.g. I-beams, connections and bracings under a bridge). The adoption of mobile sensing capabilities, through the deployment of scanners on moving vehicles on roadways and aerial pathways, facilitated many applications, such as bridge inspection, large-scale as-is modeling, VR/AR visualization and field reporting [4–7, 9, 10, 36].

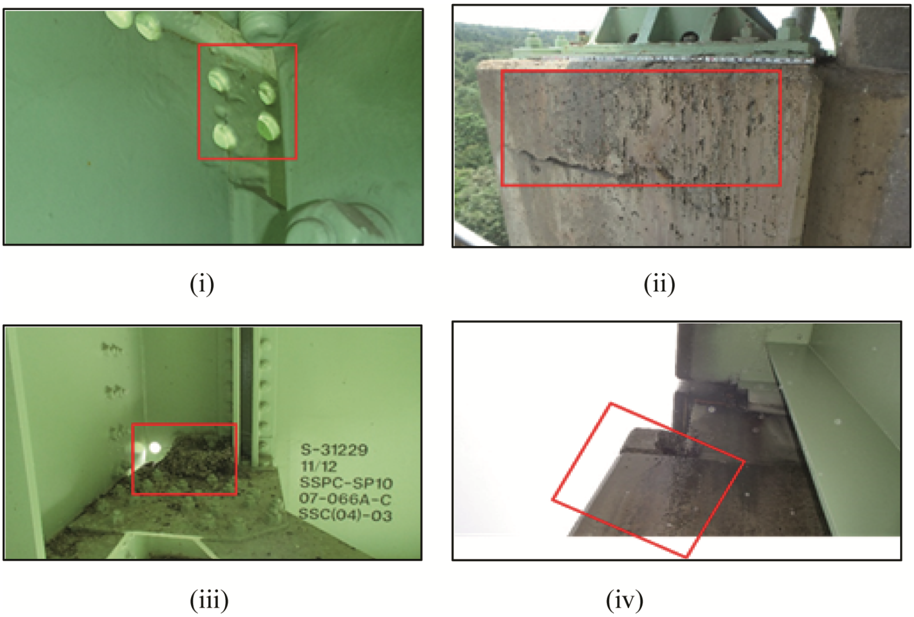
However, moving towards more mobile platforms also brings new challenges. First, mobile sensors usually capture less accurate and dense information compared to stationary sensors [98]. Theoretically, even if density can be improved by having mobile scanners do multiple passes around a scan area, it does not guarantee higher accuracy because the chances of error propagating from one pass to the other are significant [86]. Secondly, registration is no longer a static process, due to the mobility of the sensors. Instead, Simultaneous Localization and Mapping (SLAM) becomes the state-of-the-art practice. In this approach, sensors must be able to estimate its own location and scan a scene at the same time. SLAM techniques are still considered as an active research area and they also have issues like loop closure, which means that they are not always reliable when accuracy requirements are at millimeter scale [99]. Thirdly, the tradeoff between speed and density becomes critical because the point cloud captured per frame is usually sparse. A common method to get a dense output from a mobile sensor is to accumulate the data across multiple frames captured at different times. At the same time, increased data accumulation will increase the time to process the data. Therefore, developed applications are usually divided into a user-side, which enables effective interaction with users, and a server-side, which handles heavy computation to provide real-time service [100]. Finally, data from mobile sensors are not necessarily ordered and structured similar to the images captured from everyday events. Unordered data breaks temporal smoothness assumption and does not guarantee any overlapping between the data captured at different times. Since these unstructured data sets do not come from the same data collection setting, they might have high variances due to intrinsic differences of sensors or changes in environmental conditions.

4.2 From Deterministic to Probabilistic Assessment

Most of the existing approaches for data analyses employ assessment methods that are deterministic. Examples of usage scenarios that deterministic methods are applied to include: (i) Supporting visual inspection by capturing in-situ damage or collapse of a structure by creating a complete 3D record of a damaged structure and computing localization and quantification parameters of surface damage [101–103]; (ii) Creating as-is BIM by extracting crack information from laser scanners and semantically representing on structural surface by linking it to underlying BIM objects [17]; (iii) Determining bridge under clearance to maintain up-to-date and accurate structural inventory details, assess possible damage and help engineers plan for bridge improvement activities [103–105]; (iv) Measuring bridge deflection under certain conditions or usage scenarios [106]. Some research studies (using deterministic methods) have also investigated performance of laser scanners from a perspective of being able to detect specific damage types - thin crack detection [17, 107], surface flatness detection [108], structural deformations [109], and damages specific to post-earthquake [110] and tornado [111] reconnaissance efforts.

When deterministic methods are used, uncertainties associated with data capture and processing are not taken into account during analysis. For example, several types of uncertainties can occur during scanning. Typically, cases of inaccuracies in point measurements and noise are those that are influenced by factors, such as range, angle of incidence, surface reflectivity, and discontinuity edges on the objects of interest [112]. Normally, these uncertainties are of aleatory nature, and have some randomness in their values. Hence, the result of any dimensional measurement taken on the surface affected by this uncertainty cannot be regarded as exact and should be described in terms of bounds or some probabilistic distribution. However, some point cloud processing software can detect noisy points and provide an option for users to remove them from a model. In cases, where such points are removed, the nature of uncertainty can become epistemic, as the affected regions can have no scan points representing them. Otherwise, this type of uncertainty is aleatory and is calculated as the standard deviation of the distances of the points from the surface within a 2D grid cell superimposed on the BIM surface being analyzed [112]. Currently, there are no known studies that incorporate these uncertainties directly into assessment. Recent research indicates that uncertainties can impact assessment results, which in turn can have safety and economic implications [113]. Hence, there is a growing need to consider probabilistic methods, either to replace the deterministic approach, or at least to be used in conjunction with the deterministic methods wherever applicable.

While the sources of uncertainties discussed above primarily arise out of scanning related issues, such as scan quality, scan position and scan registration, it is also important to investigate some potential site conditions that contribute to similar uncertainties. These sources are illustrated in Fig. 4. Figure 4-i shows stress holes that can be mistaken by automatic computer vision algorithms for steel section loss, whereas, Fig. 4-ii shows paint freckles, which can be a potential candidate for false positive identification in crack/spall detection and quantification algorithms. Similarly, unwanted material like pigeon droppings (Fig. 4-iii), and wet conditions (Fig. 4-iv) can influence the performance of defect assessment by traditional computer vision algorithms. These situations demonstrate difficulties in identifying problematic areas due to ambiguities existing in civil engineering scenes.



(i) (ii)

(iii) (iv)

Fig. 4. Showing examples where bridge conditions can contribute to uncertainty in assessment (i) stress holes (ii) paint freckles (iii) pigeon droppings (iv) wet conditions

As visual recognition of problematic areas is a very domain-specific task even within different branches of the AEC/FM industry, it is a challenge to design suitable handcrafted features that can work well (in terms of recognition performance) specifically for these scenarios (as discussed in Sect. 2). Deep neural algorithms [114–116] provide an opportunity to bypass this step, because they automatically create features in an implicit manner over different network layers by learning the structure from the data. Even with this advantage, for probabilistic assessment using deep neural algorithms to be computationally tractable and effective, the contextual relationships between structural elements and site conditions should be modeled and represented (in a computer-interpretable form, e.g. semantically-rich 3D models).

4.3 From the Use of a Single Sensor to Sensor Fusion

Increasing number of scanning/imaging workflows adopt integrated systems instead of a single sensor. As discussed in Sect. 3, sensor fusion can be employed to address the problem of lack of 3D features when doing semantic analysis. Besides, there are three advantages of using an integrated system rather than a single sensor. First, an integrated system can mitigate several shortcomings of a single sensor environment. For example, point cloud, providing 3D spatial information, is a good complement to images that provide color and other visual information, and vice versa. Secondly, an integrated system can provide complete information of a scene, which could be helpful for further semantic analysis. For example, an integrated system that can capture images with orientation and locations can first identify workers and vehicles on an image, then track their positions using the orientation and location information, thus facilitating the application of monitoring worker activities. Finally, integrated systems can better support automation of different aspects of 3D imaging workflows. For example, with additional information such as poses of 3D imaging equipment, a system can be deployed to a robot or a drone and conduct automatic scanning of infrastructure systems [32, 117, 118].

At the same time, sensor fusion also brings many new challenges. The first challenge is cross-sensor calibration, such as aligning images to point clouds and vice versa. Though offline calibration is becoming mature, online calibration is still not robust enough hindering possible migration to ubiquitous sensing [119]. The second one is the choice of a platform. An integrated system can be a deployed as a backpack, a handheld device, a robot, or a drone, and applications drive the suitability and usability of a system in a specific environment. For example, while a drone equipped with laser scanner and camera is good for bridge inspection, a handheld device might be more suitable for indoor as-is modeling. The choice of sensor and platform with respect to different AEC/FM applications and usage scenarios still need to be researched.

5 Conclusions

In conclusion, this paper reviews primary challenges associated with utilization of 3D imaging technologies within the AEC/FM industry. The primary contribution of this paper to the current knowledge body is the assessment of existing technologies and algorithms in the context of specific challenges posed within this industry and providing future research directions.

Four categories of challenges are identified: occlusions, lack of 3D features, quality-efficiency tradeoff, and interoperability. For each category of challenges, the paper synthesizes the state-of-the-art practices on how they aim to address these challenges. The state-of-the-art practices target addressing occlusions by planning the time and path of a data capturing event to optimize coverage and visibility, by employing mobile sensors to bypass occlusions on the fly, by detecting and removing unnecessary objects in a scene through scene understanding, and by letting the user be aware of occlusions and enabling them to handle those manually. To address the need for effective 3D features, the current practices develop hand-crafted features for a specific application, fuse multiple sensors to leverage information other than 3D features, convert 3D imaging representations to 2D space and use existing 2D processing techniques, and develop algorithms to learn 3D features from labeled data. Researchers also seek to balance the tradeoff between quality and efficiency of data capture through carefully choosing a data capturing platform and improving automation of data capture. As for the efforts of improving interoperability, many industry-driven data formats and modeling techniques have been proposed to standardize the 3D imaging process.

The following three major trends are being observed in the deployment of 3D imaging within the AEC/FM industry: (1) from static to mobile scanning, (2) from deterministic to probabilistic methodologies for analyses, and (3) from the usage of a single type of sensor to multi-sensor fusion. There are also many challenges emerging with these trends such as the need for accurate SLAM methods and cross-sensor calibration and registration issues when using sensor fusion. With respect to emerging challenges, future research directions towards applying 3D imaging in AEC/FM industries could be: (1) To develop semantic SLAM algorithms that can support scene understanding while doing sensor localization and scene mapping; (2) To develop probabilistic decision-making frameworks that have 3D sensing data as inputs; (3) To develop accurate cross sensor registration methods to support sensor fusion; (4) Lastly, to establish standardized AEC/FM datasets to benchmark scene understanding approaches and support transfer learning.

Acknowledgements. The project is funded by a grant from the National Science Foundation (NSF), # 1534114. NSF's support is gratefully acknowledged. Any opinions, findings, conclusions or recommendations presented in this paper are those of authors and do not necessarily reflect the views of the NSF.

References

1. Baik, A.: From point cloud to Jeddah Heritage BIM Nasif Historical House – case study. Digit. Appl. Archaeol. Cult. Herit. **4**, 1–18 (2017)
2. Grussenmeyer, P., Al Khalil, O.: From metric image archives to point cloud reconstruction: case study of the great Mosque of Aleppo in Syria. ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-2/W5, pp. 295–301 (2017)
3. Jaklič, A., Erič, M., Mihajlović, I., Stopinšek, Ž., Solina, F.: Volumetric models from 3D point clouds: the case study of sarcophagi cargo from a 2nd/3rd century AD Roman shipwreck near Sutivan on island Brač. Croatia. J. Archaeol. Sci. **62**, 143–152 (2015)
4. Xu, Y., He, J., Tuttas, S., Stilla, U.: Reconstruction of scaffolding components from photogrammetric point clouds of a construction site. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. II-3/W5, pp. 401–408 (2015)
5. Barazzetti, L.: Parametric as-built model generation of complex shapes from point clouds. Adv. Eng. Inf. **30**, 298–311 (2016)

6. Xiong, X., Adan, A., Akinci, B., Huber, D.: Automatic creation of semantically rich 3D building models from laser scanner data. Autom. Constr. **31**, 325–337 (2013)
7. Becker, S., Peter, M., Fritsch, D.: Grammar-supported 3D indoor reconstruction from point clouds for “As-Built” Bim. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. II-3/W4, pp. 17–24 (2015)
8. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Reconstructing building interiors from images. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 80–87 (2009)
9. Díaz-Vilariño, L., Khoshelham, K., Martínez-Sánchez, J., Arias, P.: 3D modeling of building indoor spaces and closed doors from imagery and point clouds. Sensors **15**, 3491–3512 (2015)
10. Xiao, J., Furukawa, Y.: Reconstructing the world’s museums. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 668–681. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33718-5_48
11. Golparvar-Fard, M., Peña-Mora, F., Arboleda, C.A., Lee, S.: Visualization of construction progress monitoring with 4D simulation model overlaid on time-lapsed photographs. J. Comput. Civ. Eng. **23**, 391–404 (2009)
12. Turkan, Y., Bosché, F., Haas, C.T., Haas, R.: Tracking secondary and temporary concrete construction objects using 3D imaging technologies. In: Computing in Civil Engineering, pp. 749–756. American Society of Civil Engineers, Reston (2013)
13. Golparvar-Fard, M., Peña-Mora, F., Savarese, S.: Automated progress monitoring using unordered daily construction photographs and IFC-based building information models. J. Comput. Civ. Eng. **29**, 4014025 (2015)
14. Wang, J., Zhang, S., Teizer, J.: Geotechnical and safety protective equipment planning using range point cloud data and rule checking in building information modeling. Autom. Constr. **49**, 250–261 (2015)
15. Teizer, J., Caldas, C.H., Haas, C.T.: Real-time three-dimensional occupancy grid modeling for the detection and tracking of construction resources. J. Constr. Eng. Manag. **133**, 880–888 (2007)
16. Kim, H., Kim, K., Kim, H.: Data-driven scene parsing method for recognizing construction site objects in the whole image. Autom. Constr. **71**, 271–282 (2016)
17. Anil, E.B., Tang, P., Akinci, B., Huber, D.: Deviation analysis method for the assessment of the quality of the as-is Building Information Models generated from point cloud data. Autom. Constr. **35**, 507–516 (2013)
18. Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A.: Automatic reconstruction of as-built building information models from laser-scanned point clouds: a review of related techniques. Autom. Constr. **19**, 829–843 (2010)
19. Chaiyasarn, K., Kim, T.-K., Viola, F., Cipolla, R., Soga, K.: Distortion-free image mosaicing for tunnel inspection based on robust cylindrical surface estimation through structure from motion. J. Comput. Civ. Eng. **30**, 4015045 (2016)
20. Balado, J., Díaz-Vilariño, L., Arias, P., Soilán, M.: Automatic building accessibility diagnosis from point clouds. Autom. Constr. **82**, 103–111 (2017)
21. Vidas, S., Moghadam, P., Bosse, M.: 3D thermal mapping of building interiors using an RGB-D and thermal camera. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 2311–2318 (2013)
22. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from Internet photo collections. Int. J. Comput. Vis. **80**, 189–210 (2008)
23. Pătrăucean, V., Armeni, I., Nahangi, M., Yeung, J., Brilakis, I., Haas, C.: State of research in automatic as-built modelling. Adv. Eng. Inf. **29**, 162–171 (2015)

24. Seo, J., Han, S., Lee, S., Kim, H.: Computer vision techniques for construction safety and health monitoring. Adv. Eng. Inf. **29**, 239–251 (2015)
25. Yang, J., Park, M.W., Vela, P.A., Golparvar-Fard, M.: Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future. Adv. Eng. Inf. **29**, 211–224 (2015)
26. Fathi, H., Dai, F., Lourakis, M.: Automated as-built 3D reconstruction of civil infrastructure using computer vision: achievements, opportunities, and challenges. Adv. Eng. Inf. **29**, 149–161 (2015)
27. Brilakis, I., Dai, F., Radopoulou, S.-C.: Achievements and challenges in recognizing and reconstructing civil infrastructure. In: Dellaert, F., Frahm, J.-M., Pollefeys, M., Leal-Taixé, L., Rosenhahn, B. (eds.) Outdoor and Large-Scale Real-World Scene Analysis. LNCS, vol. 7474, pp. 151–176. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34091-8_7
28. Lu, Q., Lee, S.: Image-based technologies for constructing as-is building information models for existing buildings. J. Chem. Inf. Model. **31** (2017)
29. Cho, Y.K., Ham, Y., Golpavar-Fard, M.: 3D as-is building energy modeling and diagnostics: a review of the state-of-the-art. Adv. Eng. Inf. **29**, 184–195 (2015)
30. Guo, H., Yu, Y., Skitmore, M.: Visualization technology-based construction safety management: a review. Autom. Constr. **73**, 135–144 (2017)
31. Mukupa, W., Roberts, G.W., Hancock, C.M., Al-Manasir, K.: A review of the use of terrestrial laser scanning application for change detection and deformation monitoring of structures. Surv. Rev. 1–18 (2016)
32. Ham, Y., Han, K.K., Lin, J.J., Golparvar-Fard, M.: Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works. Vis. Eng. **4**, 1 (2016)
33. Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., Fieguth, P.: A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. Adv. Eng. Inf. **29**, 196–210 (2015)
34. Rabbani, T., van den Heuvel, F. a, Vosselman, G.: Segmentation of point clouds using smoothness constraint. In: Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - Comm. V Symp. 'Image Eng. Vis. Metrol. 36, 248–253 (2006)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)
36. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Reconstructing building interiors from images. In: 2009 IEEE 12th International Conference Computer Vision, pp. 80–87 (2009)
37. Armeni, I., Sener, O., Zamir, A., Jiang, H.: 3D semantic parsing of large-scale indoor spaces. CVPR, pp. 1534–1543 (2016)
38. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5122–5130 (2017)
39. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Li, F.-F.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
40. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

41. Lazebnik, S., Schmid, C., Ponce, J., Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories To cite this version: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. (2010)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neu- ral Information Processing Systems, pp. 1097–1105 (2012)
43. Lehtola, V.V., Kaartinen, H., Nüchter, A., Kaijaluoto, R., Kukko, A., Litkey, P., Honkavaara, E., Rosnell, T., Vaaja, M.T., Virtanen, J.P., Kurkela, M., El Issaoui, A., Zhu, L., Jaakkola, A., Hyyppä, J.: Comparison of the selected state-of-the-art 3D indoor scanning and point cloud generation methods. Remote Sens. **9**, 796 (2017)
44. Yang, S.W., Wang, C.C.: Dealing with laser scanner failure: mirrors and windows. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3009–3015 (2008)
45. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. IEEE Trans. Robot. **32**, 1309–1332 (2016)
46. Książek, M.V., Nowak, P.O., Kivrak, S., Rosłon, J.H., Ustinovichius, L.: Computer-aided decision-making in construction project development. J. Civ. Eng. Manag. **21**, 248–259 (2015)
47. Larsson, S., Kjellander, J.A.P.: Path planning for laser scanning with an industrial robot. Rob. Auton. Syst. **56**, 615–624 (2008)
48. Landa, Y., Tsai, R.: Visibility of point clouds and exploratory path planning in unknown environments. Commun. Math. Sci. **6**, 881–913 (2008)
49. Arora, S., Scherer, S.: Randomized algorithm for informative path planning with budget constraints. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 4997–5004 (2017)
50. Akinci, B., Boukamp, F., Gordon, C., Huber, D., Lyons, C., Park, K.: A formalism for utilization of sensor systems and integrated project models for active construction quality control. Autom. Constr. **15**, 124–138 (2006)
51. Liu, T., Carlberg, M., Chen, G., Chen, J., Kua, J., Zakhor, A.: Indoor localization and visualization using a human-operated backpack system. In: 2010 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2010 - Conference Proceedings, pp. 1–10. IEEE (2010)
52. Pu, S., Vosselman, G.: Knowledge based reconstruction of building models from terrestrial laser scanning data. ISPRS J. Photogramm. Remote Sens. **64**, 575–584 (2009)
53. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: efficient and robust 3D object recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 998–1005. IEEE (2010)
54. Diaz-Vilarino, L., Boguslawski, P., Khoshelham, K., Lorenzo, H., Mahdjoubi, L.: Indoor navigation from point clouds: 3D modelling and Obstacle Detection. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 41, pp. 275–281 (2016)
55. Kasireddy, V., Akinci, B.: Challenges in generation of as-is bridge information model: a case study. In: Proceedings of the 32nd International Symposium on Automation and Robotics in Construction and Mining: Connected to the Future (2015)
56. Velodyne LiDAR HDL-64E. http://velodynelidar.com/hdl-64e.html
57. Yoder, L., Scherer, S.: Autonomous exploration for infrastructure modeling with a micro aerial vehicle. In: Wettergreen, D.S., Barfoot, T.D. (eds.) Field and Service Robotics. STAR, vol. 113, pp. 427–440. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27702-8_28

58. Bosché, F., Ahmed, M., Turkan, Y., Haas, C.T., Haas, R.: The value of integrating Scan-to-BIM and Scan-vs-BIM techniques for construction monitoring using laser scanning and BIM: the case of cylindrical MEP components. Autom. Constr. **49**, 201–213 (2015)
59. Liu, T., Carlberg, M., Chen, G., Chen, J., Kua, J., Zakhor, A.: Indoor localization and visualization using a human-operated backpack system. In: Proceedings of the 2010 International Conference on Indoor Positioning and Indoor Navigation. IPIN 2010, pp. 15–17 (2010)
60. Metni, N., Hamel, T.: A UAV for bridge inspection: visual servoing control law with orientation limits. Autom. Constr. **17**, 3–10 (2007)
61. Maturana, D., Scherer, S.: VoxNet: A 3D convolutional neural network for real-time object recognition. Iros, pp. 922–928 (2015)
62. Armeni, I., Sener, O., Zamir, A., Jiang, H.: 3D semantic parsing of large-scale indoor spaces. In: CVPR, pp. 1534–1543 (2016)
63. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3D classification and segmentation (2016)
64. Hong, S., Jung, J., Kim, S., Cho, H., Lee, J., Heo, J.: Semi-automated approach to indoor mapping for 3D as-built building information modeling. Comput. Environ. Urban Syst. **51**, 34–46 (2015)
65. Han, K.K., Golparvar-Fard, M.: Appearance-based material classification for monitoring of operation-level construction progress using 4D BIM and site photologs. Autom. Constr. **53**, 44–57 (2015)
66. Irschara, A., Zach, C., Frahm, J.-M., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2599–2606 (2009)
67. Sattler, T., Leibe, B., Kobbelt, L.: Fast image-based localization using direct 2D-to-3D matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 667–674 (2011)
68. Dai, F., Rashidi, A., Brilakis, J., Vela, P.: Comparison of image-based and time-of-flight-based technologies for 3D reconstruction of infrastructure (2012). http://ascelibrary.org/doi/10.1061/%28ASCE%29CO.1943-7862.0000565
69. Rebolj, D., Pučko, Z., Babič, N.Č., Bizjak, M., Mongus, D.: Point cloud quality requirements for Scan-vs-BIM based automated construction progress monitoring. Autom. Constr. **84**, 323–334 (2017)
70. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors **12**, 1437–1454 (2012)
71. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 1–4. IEEE (2011)
72. Martinez, A., Fernández, E.: Learning ROS for Robotics Programming. Packt Publishing Ltd., Birmingham (2013)
73. Radu, R.: The PCD (Point Cloud Data) file format. http://pointclouds.org/documentation/tutorials/pcd_file_format.php
74. Lee, S.H., Kim, B.G.: IFC extension for road structures and digital modeling. Procedia Eng. **14**, 1037–1042 (2011)
75. Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M.: SEMANTIC3D.net: a new large-scale point cloud classification benchmark
76. Rusu, R.B., Blodow, N., Beetz, M.: Fast Point Feature Histograms (FPFH) for 3D registration. In: 2009 IEEE International Conference on Robotics and Automation, pp. 3212–3217 (2009)

77. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D recognition and pose using the viewpoint feature histogram. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2155–2162. IEEE (2010)
78. Mastin, A., Kepner, J., Fisher Iii, J.: Automatic registration of LIDAR and optical images of urban scenes. In: CVPR (2009)
79. Golparvar-Fard, M., Bohn, J., Teizer, J., Savarese, S., Peña-Mora, F.: Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. Autom. Constr. **20**, 1143–1155 (2011)
80. Park, Y., Yun, S., Won, C.S., Cho, K., Um, K., Sim, S.: Calibration between color camera and 3D LIDAR instruments with a polygonal planar board. Sensors **14**, 5333–5353 (2014)
81. Stamos, I., Allen, P.K.: Geometry and texture recovery of scenes of large scale. Comput. Vis. Image Underst. **88**, 94–118 (2002)
82. J. Huang, S.Y.: Point cloud labeling using 3D convolutional neural network. In: International Conference on Pattern Recognition, pp. 1–6 (2016)
83. Pan, S.J., Yang, Q.: A survey on transfer learning (2010). https://www.cse.ust.hk/~qyang/Docs/2009/tkde_transfer_learning.pdf
84. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3D classification and segmentation. In: CVPR (2016)
85. Yan, Y., Guldur, B., Yoder, L., Kasireddy, V., Huber, D., Scherer, S., Akinci, B., Hajjar, J.F.: Automated damage detection and structural modelling with laser scanning. In: Structural Stability Research Council Annual Stability Conference 2016, SSRC 2016 (2016)
86. Kasireddy, V., Akinci, B.: A case study on comparative analysis of 3D point clouds from UAV mounted and terrestrial scanners for bridge condition assessment. In: Proceedings of the Lean & Computing in Construction Congress (LC3). CIB W78, Heraklion, Greece (2017) (accepted)
87. Dai, F., Rashidi, A., Brilakis, J., Vela, P.: Comparison of image-based and time-of-flight-based technologies for 3D reconstruction of infrastructure. Constr. Res. Congr. **139**, 929–939 (2012)
88. Huang, J., Wang, Z., Gao, J., Huang, Y., Towers, D.P.: High-precision registration of point clouds based on sphere feature constraints. Sensors **17**, 72 (2017)
89. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, vol. 1, pp. 519–528 (2006)
90. Succar, B.: Building information modelling framework: a research and delivery foundation for industry stakeholders. Autom. Constr. **18**, 357–375 (2009)
91. Huber, D.: The ASTM E57 file format for 3D imaging data exchange. In: Three-Dimensional Imaging, Interaction, and Measurement (2011)
92. Kiziltas, S., Akinci, B., Ergen, E., Tang, P., Gordon, C.: Technological assessment and process implications of field data capture technologies for construction and facility/infrastructure management. Electron. J. Inf. Technol. Constr. **13**, 134–154 (2008)
93. Wedding, J., Probert, D.: Mastering AutoCAD Civil 3D 2009. Wiley, Chichester (2008)
94. Khemlani, L.: Autodesk Revit: implementation in practice. White Pap. Autodesk (2004)
95. Tang, P., Anil, E.B., Akinci, B., Huber, D.: Efficient and effective quality assessment of as-is building information models and 3D laser-scanned data. In: Computing in Civil Engineering, pp. 486–493 (2011)
96. Anil, E.B., Tang, P., Akinci, B., Huber, D.: Assessment of quality of as-is building information models generated from point clouds using deviation analysis. In: Environmental Engineering, vol. 7864, p. 78640F–13 (2011)

97. Autodesk Inc.: Autodesk Recap (2015)
98. Velodyne LiDAR HDL-64E. http://hypertech.co.il/wp-content/uploads/2015/12/HDL- 64E-Data-Sheet.pdf
99. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans. Robot. **31**, 1147–1163 (2015)
100. Bae, H., Golparvar-Fard, M., White, J.: Image-based localization and content authoring in structure-from-motion point cloud models for real-time field reporting applications. J. Comput. Civ. Eng. **29**, B4014008 (2015)
101. Olsen, M.J., Kuester, F., Chang, B.J., Hutchinson, T.C.: Terrestrial laser scanning-based structural damage assessment. J. Comput. Civ. Eng. **24**, 264–272 (2010)
102. Teza, G., Galgaro, A., Moro, F.: Contactless recognition of concrete surface damage from laser scanning and curvature computation. NDT E Int. **42**(4), 240–249 (2009)
103. Liu, W., Chen, S., Hauser, E.: LiDAR-based bridge structure defect detection. Exp. Tech. **35**, 27–34 (2011)
104. Tang, P., Akinci, B.: Automatic execution of workflows on laser-scanned data for extracting bridge surveying goals. Adv. Eng. Inf. **26**, 889–903 (2012)
105. Tang, P., Chen, G., Shen, Z., Ganapathy, R.: A spatial-context-based approach for automated spatial change analysis of piece-wise linear building elements. Comput. Civ. Infrastruct. Eng. **31**, 65–80 (2016)
106. Chen, S.: Laser Scanning Technology for Bridge Monitoring. InTech (2012)
107. Laefer, D.F., Truong-Hong, L., Carr, H., Singh, M.: Crack detection limits in unit based masonry with terrestrial laser scanning. NDT E Int. **62**, 66–76 (2014)
108. Loprencipe, G., Cantisani, G.: Evaluation methods for improving surface geometry of concrete floors: a case study. Case Stud. Struct. Eng. **4**, 14–25 (2015)
109. Kayen, R., Pack, R.T., Bay, J., Sugimoto, S., Tanaka, H.: Terrestrial-LIDAR visualization of surface and structural deformations of the 2004 Niigata Ken Chuetsu, Japan, earthquake. Earthq. Spectra. **22**, 147–162 (2006)
110. Olsen, M.J., Kayen, R.: Post-Earthquake and Tsunami 3D laser scanning forensic investigations. Forensic Eng. **2012**, 477–486 (2012)
111. Kashani, A., Crawford, P.: Automated tornado damage assessment and wind speed estimation based on terrestrial laser scanning. J. Comput. Civ. Eng. **29**, 1–10 (2014)
112. Anil, E.B., Akinci, B., Huber, D.: Representation requirements of as-is building information models generated from laser scanned point cloud data. In: Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC), Seoul, Korea (2011)
113. Kasireddy, V., Akinci, B.: Towards the integration of inspection data with bridge information models to support visual condition assessment. In: Proceedings of the Congress on Computing in Civil Engineering, pp. 644–651. American Society of Civil Engineers, Reston (2015)
114. Sermanet, P., LeCun, Y.: Traffic sign recognition with multi-scale convolutional networks. In: Neural Networks (IJCNN) (2011)
115. Protopapadakis, E., Doulamis, N.: Image based approaches for tunnels' defects recognition via robotic inspectors. In: Bebis, G., et al. (eds.) ISVC 2015. LNCS, vol. 9474, pp. 706–716. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27857-5_63
116. Maeda, H., Sekimoto, Y., Seto, T.: An easy infrastructure management method using on-board smartphone images and citizen reports by deep neural network. In: IoT in Urban Space, pp. 111–113. ACM Press, New York (2016)
117. Bang, S., Kim, H., Kim, H.: UAV-based automatic generation of high-resolution panorama at a construction site with a focus on preprocessing for image stitching. Autom. Constr. **84**, 70–80 (2017)

118. Malihi, S., Valadan Zoej, M.J., Hahn, M., Mokhtarzade, M., Arefi, H.: 3D building reconstruction using dense photogrammetric point cloud. ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLI-B3, pp. 71–74 (2016)
119. Zhao, K., Iurgel, U., Meuter, M., Pauli, J.: An automatic online camera calibration system for vehicular applications. In: 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014, pp. 1490–1492. IEEE (2014)

IoT - An Opportunity for Autonomous BIM Reconstruction?

Robert Irmler, Steffen Franz, and Uwe Rüppel

Institute of Numerical Methods and Informatics in Engineering,
Technische Universität Darmstadt,
Franziska-Braun-Str. 7, 64287 Darmstadt, Germany
irmler@iib.tu-darmstadt.de

Abstract. While the concepts and technologies of Building Information Modeling (BIM) are on their way to become an industry standard, there is an increasing demand for the digitalization of the existing building stock. Based on our previous work on this subject we present a framework for autonomous BIM reconstruction using an IoT-based approach. We show that it is possible to integrated depth sensing technologies with common devices that move regularly throughout a building. By autonomously conducting several scans of the same region it is possible to enhance the quality of the reconstructed model while reducing the human workload. The use of (semi-)autonomous devices however comes with a trade-off and in this study could not achieve results equal to human-made scans.

Keywords: BIM · Reconstruction · IoT · Depth sensor · Google Tango

1 Introduction

Building Information Models (BIM) provide a great source for computational analysis, optimization and building operation. In recent years, a great share of the Architecture, Engineering and Construction (AEC) industry worldwide has adopted the concepts and technologies of Building Information Modeling. Regarding the planning and construction phase of new building projects, these concepts are on their way to become an industry standard.

At the same time, there is an increasing demand for the digitalization of the existing building stock. The necessity for the digital reconstruction of cities and buildings arises from various application areas e.g., urban planning, pollution forecasting or disaster management. The 3D reconstruction of semantically rich building models introduces a new scope of use cases ranging from the preservation of architectural knowledge [1] to the forensic documentation of crime scenes [2]. Over the past two decades, a lot of effort and research went into exploring new technologies for the digital reconstruction of our environment. In this context, the robust reconstruction of complete indoor scenes represents a particularly challenging problem.

At the EG-ICE 2017, we presented a framework for collaborative BIM reconstruction based on handheld consumer depth cameras (Google Tango) [3]. Our framework

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 61–81, 2018.
https://doi.org/10.1007/978-3-319-91635-4_4

allows us to collaboratively scan complex indoor environments and automatically generate basic floor plans and IFC models that include rooms, walls, doors and windows. Even though this approach is feasible, it still relies on humans to conduct the scanning process.

A parallel development, which has become progressively important in recent years, is the equipment of buildings with increasingly more intelligent devices. Common objects like smoke detectors, thermostats, light bulbs, hoovers, etc. are equipped with interfaces to share their data and communicate their status. They are part of the so-called Internet of Things (IoT). Decreasing costs for sensors and further miniaturization of hardware fuel this development. Experts estimate that there will be about 30 billion of these devices by 2020 [4]. Household IoT appliances like hoover robots are already capable of scanning their surroundings to optimize their path finding. Due to decreasing hardware costs, other conventional equipment and machines like cleaning trolleys, sweeper-scrubber machines or maintenance kits can easily be modified to serve as IoT devices. Especially in office buildings, these devices are permanently present and therefore offer the opportunity to be used as drones for an autonomous BIM reconstruction.

In this paper, we present an extension of our collaborative BIM reconstruction framework [3]. Our goal is to optimize the data collection process by integrating IoT devices that are already capable of scanning their environment and extending ordinary devices with these functionalities. We especially target devices that are part of regular operations like cleaning or maintenance equipment. We will show that this IoT-based approach has the potential to make the data collection process more autonomous and can further help to address typical issues of indoor scanning like clutter and occlusion. The hypothesis here is that by scanning the same area multiple times under various conditions, errors induced by clutter and occlusion can be minimized. Therefore, we aim to use recurring activities such as cleaning or maintenance to capture multiple scans of the same area.

This paper consists of six sections including the introduction. Section 2 provides a brief review of methods targeting the reconstruction of ‘as-built’ building models. Section 3 gives an overview of current hardware frameworks and IoT devices that are capable of creating a 3D model of their environment. In Sect. 4 we present our concept for an IoT-based framework for BIM reconstruction and summarize the underlying reconstruction process. Section 5 describes the hardware setup and methodology of our case study, including the discussion of our results. Finally, Sect. 6 summarizes conclusions and outlines further research topics.

2 Related Work

Over the past two decades the technologies for the digital reconstruction of our environment have advanced rapidly. This section provides a brief overview of different approaches on the reconstruction of building structures based on sensor data.

Spatial data at city scale is usually collected via airborne LiDAR systems [5–7]. To improve data quality, the fusion of different data sources like images and GIS data has been investigated by [8]. For the envelope reconstruction of single buildings, stationary laser scanners [9] and image-based techniques [10–12] have been proven to be suitable. The robust reconstruction of complete indoor scenes however remains a particularly challenging problem. Indoor environments often exhibit high levels of clutter and occlusion which can lead to noise, artifacts and missing data. Reliable indoor reconstruction methods therefore have to be tolerant to missing data and occlusion as well as to transparent or highly reflecting surfaces like windows and mirrors. Prior work on this subject shows a great variety of different approaches each of which displays strengths and weaknesses. Many indoor reconstruction systems require substantial hardware setups [13–16], rely on user interaction [2, 17] or only generate meshed geometries, i.e. do not identify individual building elements, and therefore lack semantic depth [18, 19]. An example for a particularly challenging problem is the detection of windows within the scan data. Because windows are transparent they are shapeless in 3D space and lack discriminative power such as color and texture for image feature descriptors. Approaches on the detection and reconstruction of windows include computer vision and machine learning [16, 20, 21].

In the past, many of the approaches mentioned above have been based on stationary laser scanning or image processing methods e.g., [13, 22–24]. With the increasing availability of consumer depth cameras in recent years, an additional research and application field emerged. Consumer depth cameras like the Microsoft Kinect and the Google Tango provide a cost-effective way to gather 3D measurements of the environment by globally registering a series of images along a complex trajectory [2, 25, 26]. However, even though such devices can significantly speed up the scanning process they do not themselves solve the underlying problems to extract higher-level models from the captured point cloud data. Instead, the reconstruction of semantically enriched building models still requires a dedicated post processing of the raw scan data.

Previous work on the reconstruction of indoor environments based on consumer depth cameras focused mainly on achieving high accuracy of scan data and meshed geometries [18, 19] rather than detecting specific building structures like rooms, walls, door and windows. Moreover, the optimization of scanning time, hardware costs and computation resources are often of subordinate importance. Only little research has been conducted regarding the speed-up of the data collection and reconstruction process by using a collaborative approach. In [3] we proposed a framework for the on-site, real-time, collaborative reconstruction of multi-room indoor scenes using Google Tango-enabled devices. The collaborative approach makes the system especially suitable for use-cases that are time-critical and require concurrent collection and processing of data. Because the framework is able to process the input from multiple scanners simultaneously, the on-site data collection workload can be split among multiple users. Nevertheless, because of the common challenges associated with the scanning of indoor environments e.g., clutter and occlusion, the data collection still remains a time consuming task. Sometimes it may be necessary to perform multiple scans of the same region under varying conditions until the quality of the scan data is sufficient. We therefore consider it propitious to investigate the potential for further automation of the scanning process.

Following the thought of automating the scanning process naturally leads to the field of robotics. Similar to civil engineering, current research in robotics too targets the reconstruction of indoor environments to create navigation maps for robots in order to increase their performance and open up new application areas [27]. Additionally, due to decreasing hardware costs and the miniaturization of computers and sensors, the so-called Internet of Things (IoT) has been growing rapidly in recent years. In combination with the aforementioned research developments, household IoT appliances like cleaning robots are now capable of mapping their surroundings by using laser range finders or bumper sensors [28, 29]. This development represents a tremendous potential that can eventually be exploited to make indoor reconstruction applications available on a large scale.

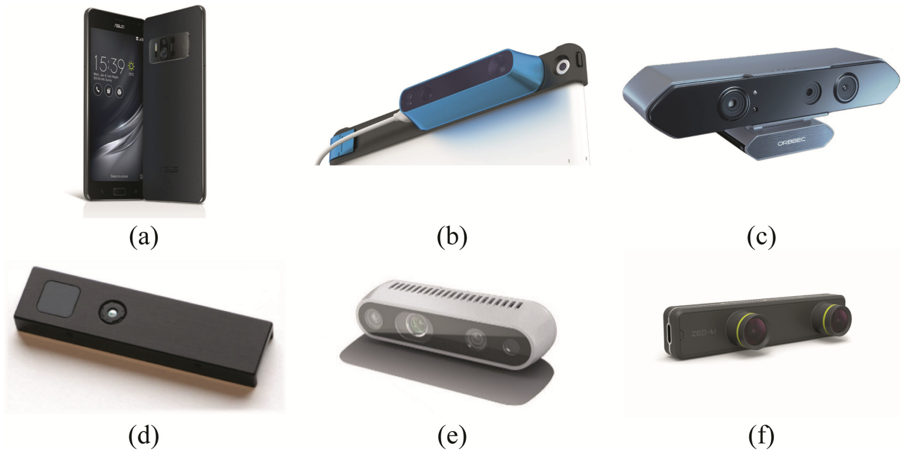
3 Mobile Technologies for Environmental Perception

This section gives a short overview of common depth sensing technologies and outlines the utilization of cleaning robots as autonomous scanning devices.

3.1 Overview of Depth Sensing Technologies and Devices

In November 2010, Microsoft released the Kinect for XBox 360, which marked the first major introduction of a depth sensor for the end-user market. The depth sensing capability of the Kinect V1 was based on an infrared sensor by Israeli developer PrimeSense in combination with other specialized hardware and software. The Kinect was primarily used as a motion tracking device and was set to broaden the XBox 360's audience by enabling the users to control and interact with their console without the need for a game controller. The Kinect received generally positive reviews from users and critics and sold over 35 million times. By the time it was finally discontinued in October 2017, the Kinect had attracted the interest of many researchers. The consumer-grade range sensor was immediately recognized as an attractive alternative to expensive laser scanners in application areas such as indoor mapping, surveillance, robotics and forensics [30] but was also investigated for a variety of other application areas such as physical rehabilitation [31], health monitoring [32], sign language recognition [33] or altitude control of quadrotor helicopters [34]. Today, just seven years after the launch of Kinect, we not only find a broad range of consumer depth camera research and applications, but also a variety of devices and technologies (see Fig. 1). There are currently more than a dozen different depth sensors or depth sensing enabled devices available.

Most of the currently available devices use an infrared sensor like the Infineon 3D Image Sensor [35] to measure distances based on the 3D Time-of-Flight (ToF) technology. A 3D ToF camera works by illuminating a scene with modulated light and observing the reflection. The phase shift between illumination and the reflection is measured and translated to distance. Another approach to calculate distances is the use of stereo photography. The ZED camera by Stereolabs for example, uses a stereo camera with two lenses to simulate binocular vision. The depth perception arises from the 'disparity' of a given point in the right and left image and can be calculated via triangulation. Regarding the utilization of depth sensing devices, there are two main groups of devices that can be identified. The first group combines sensors and special vision processors to measure and calculate real-time depth images that can be accessed through hardware interfaces. These devices typically need to be mounted on a computing device to allow for user interaction and control over the measurement process. The second group integrates the depth sensing technology directly within a computing device, like for example the Asus ZenFone AR or the Intel RealSense Robotik Development Kit.


(a) (b) (c)

(d) (e) (f)

Fig. 1. Examples of currently available depth sensors: (a) Asus ZenFone AR [36] (b) Occipital Structure [37], (c) Orbeec Persee [38], (d) Pico Flexx [39], (e) Intel RealSense D435 [40], (f) ZED Mini [41].

3.2 Cleaning Robots

Cleaning robots are popular tools and were among the first types of robots that consumers included into their daily lives. These robots rely on sensor data in order to execute their cleaning routines. The simplest type of sensors used are bumper sensors that are triggered when the robot encounters an obstacle and causes it to change course. In order to optimize operation time and to enhance their performance, some more sophisticated models use visual tracking or even laser range sensors to retrieve a better perception of their

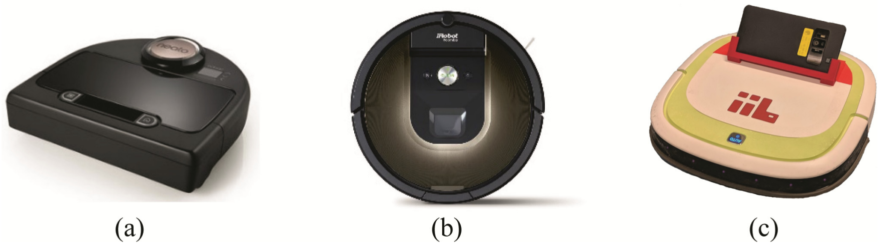

(a) (b) (c)

Fig. 2. Cleaning robots with different technologies for environmental perception, showing (a) the Neato Botvac Connected with LIDAR sensor [Photo: Neato], (b) the iRobot Roomba 980 using a visual camera and VSLAM [Photo: iRobot] and (c) the Ecovacs Deebot Slim with a mounted Asus ZenFone AR (more details in Sect. 5).

environment (see Fig. 2a/b). The sensor data is usually used internally by the robot to perform some sort of room segmentation and optimize its path-finding. If given access to the sensors directly, it is possible to use these kind of data sets as source for a (BIM) reconstruction procedure. As cleaning robots move around a building on a regular basis, they form a suitable set of devices for an IoT-based approach to building reconstruction. But even if there is no direct access to the sensor data, given the wide array of depth sensors outlined in the previous section, it is relatively easy to extend a cleaning robot with depth sensing capabilities and additional computing power (see Fig. 2c) to make use of its autonomous movements.

4 IoT-Based BIM Reconstruction

As outlined in Sect. 3, modern sensors like the Intel RealSense DepthCamera [40], the Occipital Structure Sensor [37] or the ZED Stereo Camera [41] provide the opportunity to equip devices with the ability to see, understand, and learn from their environment. The combination of multi-modal data acquisition capabilities and mobile computing power makes these systems particularly suitable for tasks where mobile on-site processing of scan data is required. By connecting these devices to the world wide web, it is further possible to combine and share information in real-time. Mobile devices like the Asus ZenFone AR already include sensors for depth perception and visual odometry and therefore provide the opportunity to collect and process scan data on a single mobile device. Additionally, intelligent household appliances like cleaning robots use sensors to map their environment in order to increase their performance e.g., [28, 29]. In the following subsections we describe our approach on using these technologies to collect and process scan data for the reconstruction of floor plans and interior building structures.

4.1 Framework for IoT-Based BIM Reconstruction

Figure 3 shows the framework for an IoT-based BIM Reconstruction. The idea is to reduce the human workload involved with the data collection process by making use of specialized IoT devices that are already present inside a building. In particular, we target mobile devices that are part of daily routines and regularly move through the building. By collecting multiple scans of the same region under different conditions, we aim to reduce the impact of errors induced by clutter, occlusion or varying lighting conditions.

IoT Cloud. The IoT Cloud comprises all scanning devices that are accessible to the server. For our particular use-case, these devices have to be able to capture and communicate scan data of their surroundings i.e., point cloud data, as well as information about their current state and pose. This can either be accomplished by making use of already integrated sensors and available APIs or by equipping the targeted appliances with a dedicated sensor and a computing device e.g., a single plate computer or smart phone, to communicate with the server. Details on the particular hardware setup utilized in this study are given in Sect. 5.

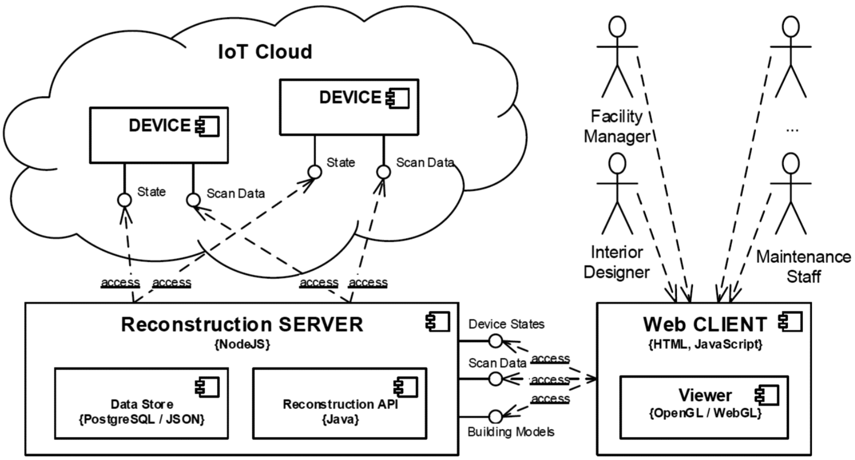


Fig. 3. Framework architecture

Reconstruction Server. The core of the framework is a NodeJS server application that provides static content (web sites), manages all client connections and acts as the central data receiver and transmitter. Data received from the IoT Cloud is cached in a data pool, processed by a dedicated reconstruction pipeline and distributed to all requesting clients. The server framework includes a data store that holds general project information, session data and geospatial building models e.g., CityGML-LoD1/LoD2. Geospatial building models like CityGML-LOD2 provide the envelope geometry of a building and can later be merged with the reconstructed interior building structures.

Web Client. The web client is a web-frontend that includes a WebGL viewer to either visualize the aggregated raw scan data or the reconstructed building model (Industry Foundation Classes - IFC). The web client only requires a web browser that supports WebGL. The client application can therefore be used on most computers, laptops and mobile phones and might be integrated with other applications depending on the specific use-case e.g., facility management, maintenance, interior design software.

4.2 Reconstruction Pipeline

Figure 4 gives an overview of our post-processing pipeline for the reconstruction of building structures i.e., rooms, walls, doors and windows. Previous work on this pipeline has been published in [3].

The pipeline is based on the processing of three distinct datasets: 1. A list of 2-dimensional space polygons; 2. A list of device poses i.e., the trajectory of the scanning device; 3. The 3D point cloud data of the recorded scene. It should be noted, that the first two datasets are already sufficient for the reconstruction of rooms, walls and doors, whereas the third dataset (3D point cloud) is only necessary for the detection of window openings and can be omitted if not available. The focus on 2-dimensional datasets for the reconstruction of rooms, walls and doors has two reasons. On the one hand it reduces

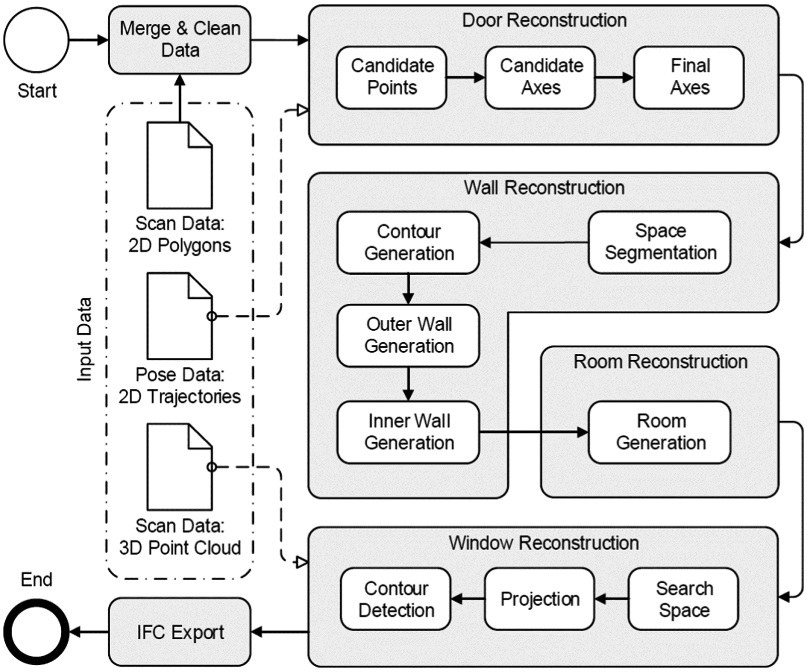


Fig. 4. Reconstruction pipeline

computational costs and communication load, and on the other hand this kind of data is congruent with the mapping data produced by cleaning robots (see [28]) which we see as a potentially major target group of devices.

The following paragraphs describe the different steps of the pipeline, ranging from the raw scan data to a reconstructed interior building model.

Merge and Clean Data. Scan data usually consist of the desired information and undesired artifacts. Artifacts usually originate from measurement errors. For example, 3D ToF depth cameras are vulnerable to transparent or highly reflecting surfaces like windows, picture frames or mirrors. Falsely recorded polygons are usually small and can be eliminated by a threshold for a minimum size of a space. Larger incorrect spaces can be eliminated by using topology constraints i.e., if a space has not been traversed and at the same time is not located within larger space, it will be classified as an error.

If there are multiple scans of the same region available, we generate a *coverage heat map* and determine how often an area has been covered by a scan. We then eliminate all data below a certain coverage threshold.

If a collaborative scan has been performed and there are datasets of multiple devices available, we merge these datasets into a single one. Figure 5 shows an example dataset before and after cleaning.

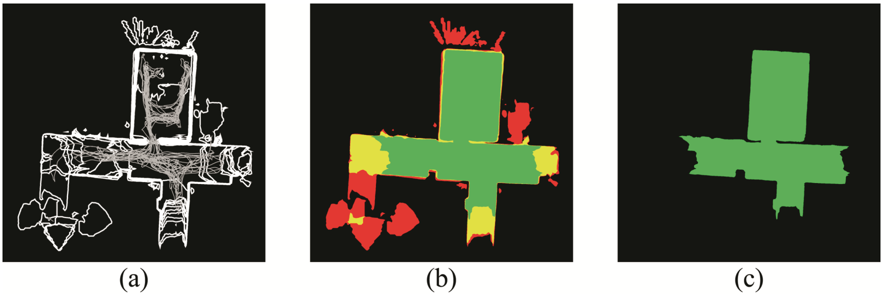


Fig. 5. Example dataset: showing (a) a set of raw input data (space polygons & trajectory) consisting of 12 individual scans, (b) the generated coverage heat map and (c) the cleaned data.

Door Reconstruction. Merging and cleaning the input data usually results in a single space polygon covering the whole scene. To detect possible door openings within this scene we look for narrow regions along the trajectory. The door reconstruction process consists of three major steps.

1. Candidate Points: We determine points of the space polygon that are close to the trajectory.
2. Candidate Axes: By fully interconnecting the previously collected ‘Candidate Points’ we retrieve a set of all connections between the Candidate Points. We then use geometric and topological conditions to either falsify each connection or classify it as a ‘Candidate Axis’ e.g., a connection gets falsified if it crosses a space boundary.
3. Final Axes: To determine the final set of door axes we divide the set of Candidate Axes into several clusters by grouping neighboring Candidate Axes together. For each cluster of Candidate Axes we determine the most likely door axis with respect to a standard door width.

Figure 6a visualizes the door reconstruction process.

Wall Reconstruction. The wall reconstruction starts with the segmentation of the cleaned space polygon. We use the previously detected door axes to split the ‘overall’ space polygon into several spaces. The next step is the construction of an outer contour of the scene. Because scan data often shows low precession in occluded or not fully coverable areas, at this stage the boundaries of the segmented spaces are usually still quite blurry and uneven. Before this data can be used for the contour generation it needs to be smoothed. Therefore, we apply the Visvalingam-Whyatt algorithm [42] on each polygon to achieve a simplified space boundary while retaining sufficient geometric detail. Based on the segmented and simplified space polygons we now generate an outer contour of the scene as a prerequisite for the following wall reconstruction. The contour is generated by applying several buffers and unification operations on the space polygons. Based on the generated contour we now begin to construct outer wall segments for each edge of the contour. Because the characteristics of the outer wall cannot be reasoned based on the available scan data we assume a standard wall thickness. The inner walls of the scene are then reconstructed based the previously generated contour and the segmented space polygons. After computing the difference between the contour polygon and each space polygon the resulting shapes are treated as the inner walls of scene. To achieve more regular shaped wall segments, we again use the Visvalingam-Whyatt algorithm to simplify their boundaries. Figure 6b–e visualize the described process.

Room Reconstruction. Due to the simplification of the contour (i.e., the outer walls) and the inner walls, the shape of the segmented spaces that were created in step 3 of the pipeline are not valid anymore. Because the boundary of a space is determined by its bounding walls, we use the now available inner wall segments to reconstruct the final shape of the spaces by computing the difference between the contour polygon and each inner wall segment (see Fig. 6f).

Window Reconstruction. In contrast to the preceding pipeline steps, windows cannot be detected by solely evaluating the 2-dimensional scan data. Instead, we now need to investigate the point cloud data. Our approach on window detection is currently limited to the peculiarities of the scanning data retrieved from a 3D ToF camera. The 3D ToF camera cannot detect transparent surfaces which manifests in a lack of point cloud data in such areas. We use this knowledge to search for 'holes' inside the point cloud data and try to determine if such a hole might represent a window. To reduce the computational costs, we use the results of the previous pipeline steps to narrow down the search area. Figure 7 illustrates the three basic steps of the window reconstruction process.

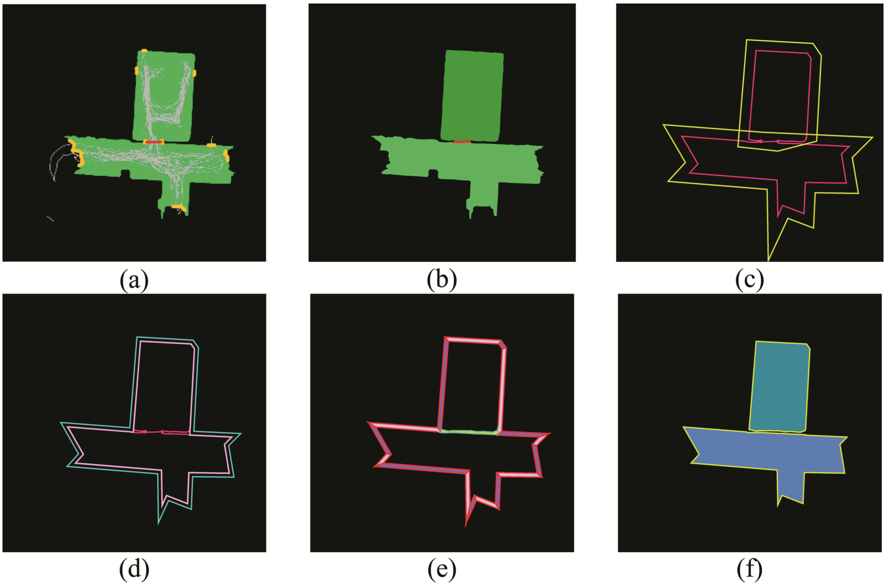


Fig. 6. Example dataset: showing (a) the door reconstruction process, (b) the segmented space polygon, (c) the simplified and buffered polygons, (d) the determined outer contour of the scene, (e) the segmented outer and inner walls and (f) the reconstructed rooms based on the previously detected doors and walls.

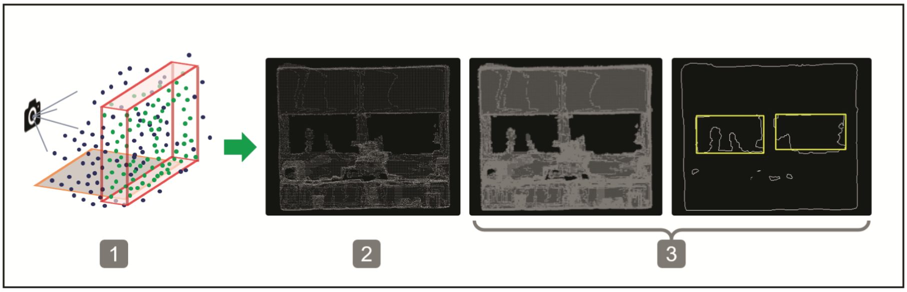


Fig. 7. Example dataset: illustrating the process of the window detection with (1) showing the creation of the search area and search space, (2) showing a binary image resulting from the projection of the points inside the search space and (3) showing image processing steps of the window contour detection.

1. Search Space: We begin by generating a 2-dimensional search area along each line segment of the previously created contour of the scene. By extruding the area along the z-axis we retrieve a 3-dimensional search space.
2. Projection: Using the search space we create a projection of all points inside the search space onto a plane corresponding to the contour edge. This gives us a binary image, where each point inside the search space is represented by a white dot.
3. Contour Detection: Before we can use a contour detection algorithm to recognize 'data-holes' i.e., black areas in the image that might be windows, we first have to apply blurring filters (dilation, gauss) on the image. This is necessary because otherwise every gap between two points would qualify as a 'data-hole', thus corrupting the detection result. After applying the blur filters we achieve a gray-scale image which we can use to detect contours that qualify as windows by using standard OpenCV procedures.

IFC Export. Figure 8 shows the resulting floor plan and the generated IFC model of an example dataset. The export is completely automated and does not require any manual steps.

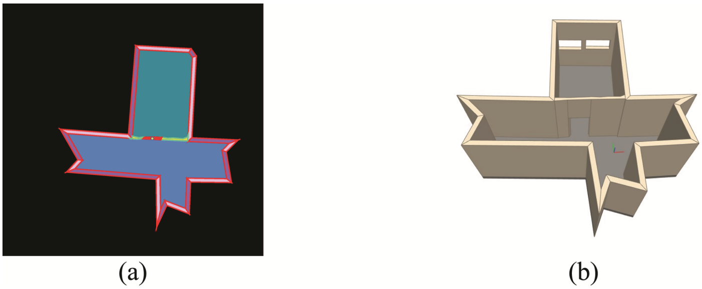


Fig. 8. Example dataset: showing (a) the resulting floor plan consisting of rooms, walls, doors and windows and (b) the automatically generated IFC model.

5 Case Study

In this section we present an implementation of the framework described in the previous section. We start by explaining the hardware setup used and discuss the software architecture and the workflow. We then describe the test sites where we evaluated the system and present our results.

5.1 Hardware Setup and Framework Implementation

For this case study, we developed an Android application based on the Google Tango SDK. The application runs on an Asus ZenFone AR, which acts as our sensor platform. The application controls the scanning process and handles the communication with the reconstruction server. Figure 9 shows the implemented framework. We chose to investigate two different application scenarios for IoT-based BIM reconstruction. In the first scenario we mounted the ZenFone onto a cleaning robot. This scenario represents the use of IoT-based BIM reconstruction in context with automated building operations using autonomous, self-moving devices that do not require permanent human interaction. The second scenario represents the extension of manual operations with automated scanning devices. Therefore, we equipped an ordinary industrial floor cleaning machine with a mount for the ZenFone. Figure 10 shows the two hardware setups.

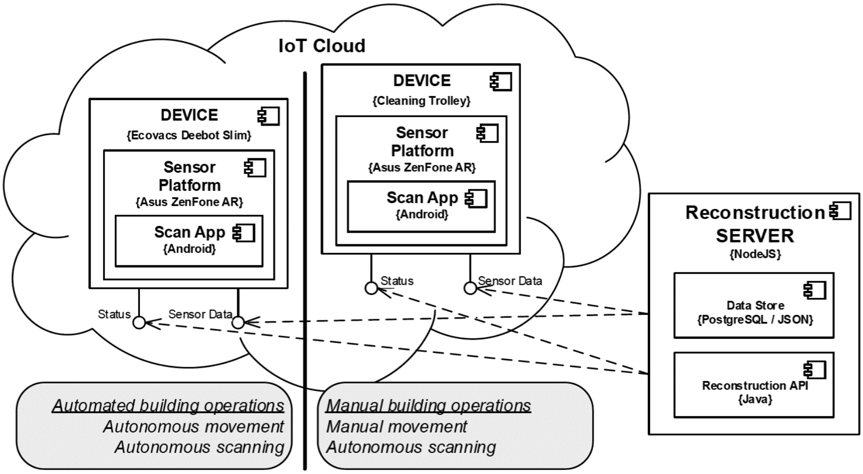


Fig. 9. Implemented framework architecture.

The sensor on the cleaning robot is positioned 80 mm above ground and tilted by an angle of 45° compared to the vertical axis. On the cleaning machine, the sensor is positioned 960 mm above ground and also tilted by an angle of 45°. Due to the different height of the sensor position, the two setups lead to slightly different perceptions of the environment and therefore to different scanning data. More discussion on this topic is provided in Subsect. 5.3.

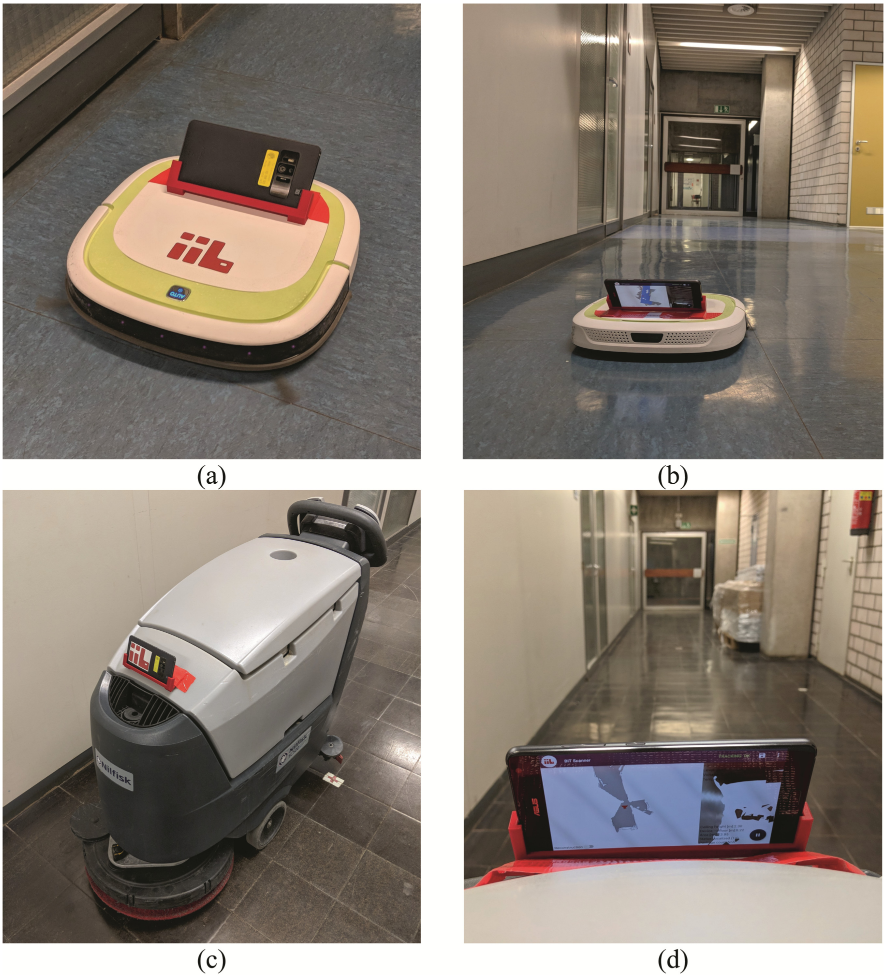


Fig. 10. Hardware setups of the case study, showing an Asus ZenFone AR mounted on (a)/(b) a cleaning robot and (c)/(d) on a floor cleaning machine

Given the two hardware setups, we implemented the following workflow as visualized in Fig. 11. The scanning process begins by starting the device and the scan app. The app then searches the server for an Area Description File (ADF) which is a special file format used by the Tango API to store a generated area description. The ADF enables the device to recognize its position within a known environment and to correct drift errors. The ADF also allows for the correct alignment of multiple scans by acting as the reference coordinate system. If no ADF is available, the app enables the learning mode of the Tango API to generate an area description along the way. Either way, after the device has localized itself within its known area, it keeps tracking its position using visual-inertial odometry (VIO). At the same time, the device is collecting data from the depth sensor and saves it along with other status and trajectory information. If a WiFi connection is available, the scan data is continuously posted to the server where it is stored in a database and can later be handed over to the reconstruction pipeline. If the device loses track of its position, it pauses the scanning process until it is re-localized.

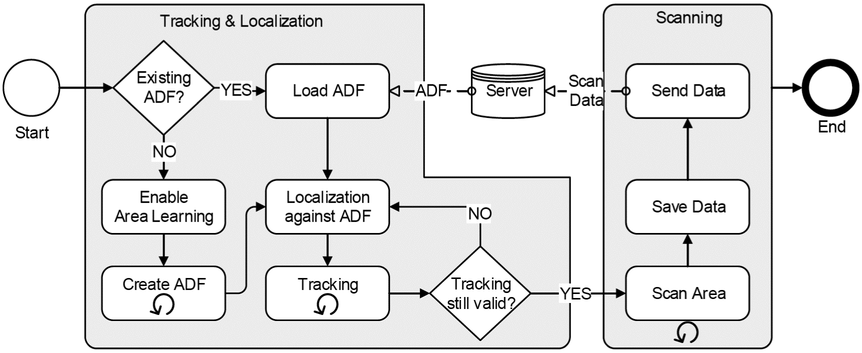


Fig. 11. Implemented workflow based on Google Tango SDK

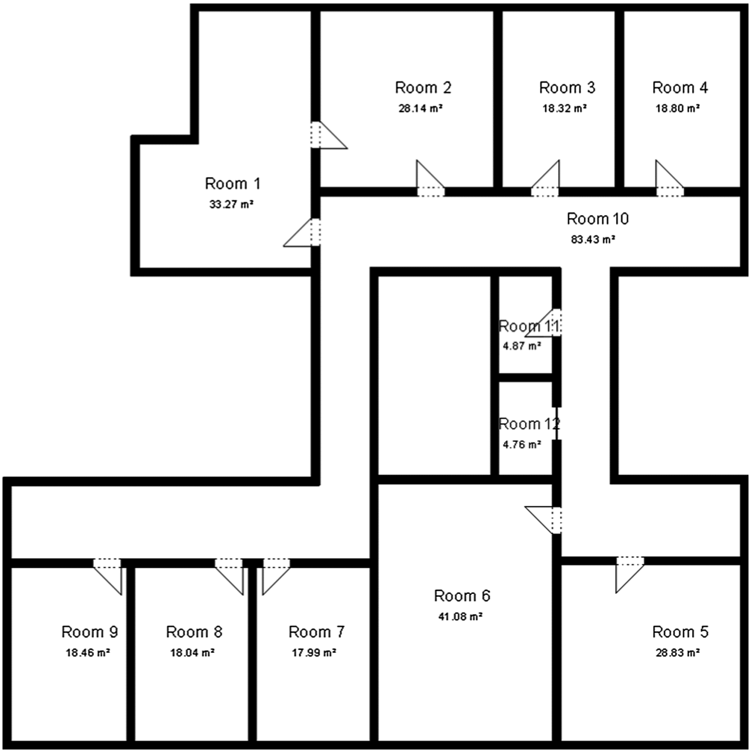


Fig. 12. Test site “Office”

5.2 Test Site and Methodology

The implemented framework and hardware setups were evaluated using the test environment shown in Fig. 12. The test site is a floor in an office building built in the 1960s that offers a variety of common challenges for indoor reconstruction like occluded areas and reflecting or transparent surfaces.

Table 1 lists our test scenarios. In the first scenario (“CM”) we used the cleaning machine to conduct the scanning process. Because such machines are usually only used in larger areas, we chose to only scan the hallway to fit the targeted application scenario. In the second scenario (“CR”) we used the cleaning robot to autonomously perform a multi-room scan between rooms 3 and 4 (see Fig. 12). The scenarios “H-1” and “H-2” serve as our comparative datasets. These datasets originate from scans conducted by a human.

Table 1. Test scenarios

Name	Hardware setup	Scanned rooms	Number of scans
CM	Cleaning Machine + Asus ZenFone AR	10	6
CR	Cleaning Robot + Asus ZenFone AR	3 + 4 + 10*	6
H-1	Human + Asus ZenFone AR	10	1
H-2	Human + Asus ZenFone AR	3 + 4 + 10*	1

*: partially

To investigate our hypothesis, that multiple scans of the same region may increase data quality and reconstruction results, we performed 6 individual scans for the scenarios “CM” and “CR”. For each scenario we then overlaid the datasets to compute a *coverage heat map* to determine how often each part of a scanned area has been covered. The underlying assumption here is, that ‘real’ boundaries and objects will be recognized most of the time, whereas errors may occur more random. Using the coverage heat map it is then possible to highlight areas that have been recognized sufficiently often and therefore help to separate correct data from errors.

5.3 Results and Discussion

Figure 13 visualizes the results of test scenario “CM”. This scenario simulated the use of a cleaning machine as a drone for the scanning process. We can see that most of the 6 individual scans achieved good results. In some cases however, we can also observe some drift errors as shown for example in Fig. 13b. The use of a coverage heat map (see Fig. 13g) proved to be suitable to eliminate most of such errors. The resulting floor plan is close to the quality achieved by a human-made scan. It should be noted however, that this scenario represents a fairly simple one because the scanned hallway does not show a lot of clutter or occlusion and does not include windows or doors that need to be detected.

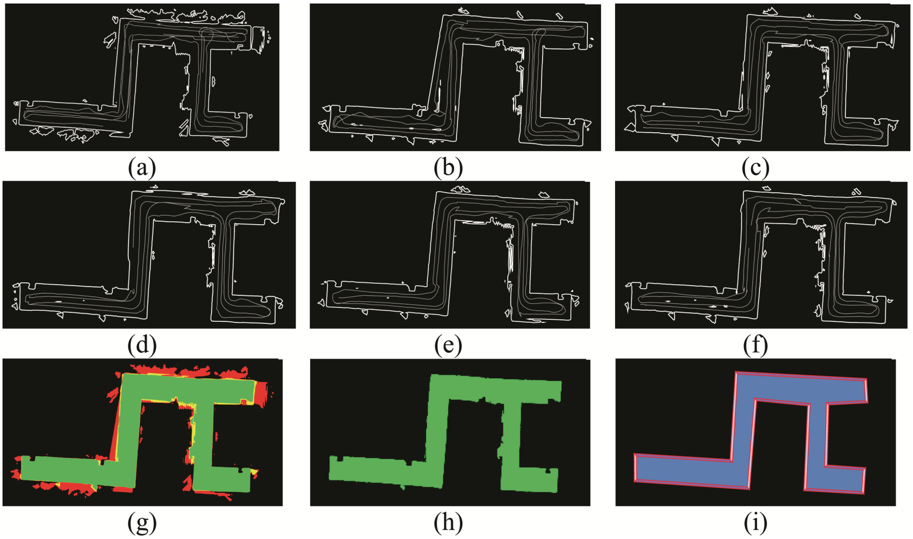


Fig. 13. Results of test scenario “CM”, showing (a) – (f) the raw scan data of 6 individual scans, (g) the coverage heat map, (h) the cleaned polygon and (i) the reconstructed floor plan.

Figure 14 shows the results of the test scenario “CR” which simulated the use of a cleaning robot as autonomous drone for the scanning process. In this scenario too, 6 individual scans form the basis for the reconstruction process. In contrast to the previous scenario, a multi-room scan was performed covering two offices connected by a hallway. The results show that the recorded scan data are by far not as coherent as in the previous scenario. One reason for that is suspected to be the different positioning of the sensor. Because of the very low position of the depth sensor and camera, the cleaning robot has a limited field of view. Due to the cluttered environment and because the robot is able to move below furniture objects like tables and chairs we often experienced drift errors when the Tango device lost its visual tracking. Because of the many drift errors and artifacts in the scan data the coverage analysis was not able to produce a reasonable input dataset for the reconstruction pipeline. The reconstructed floor plan is only a rough approximation of reality. More complex objects like doors or windows could not be detected based on the quality of the scan data.

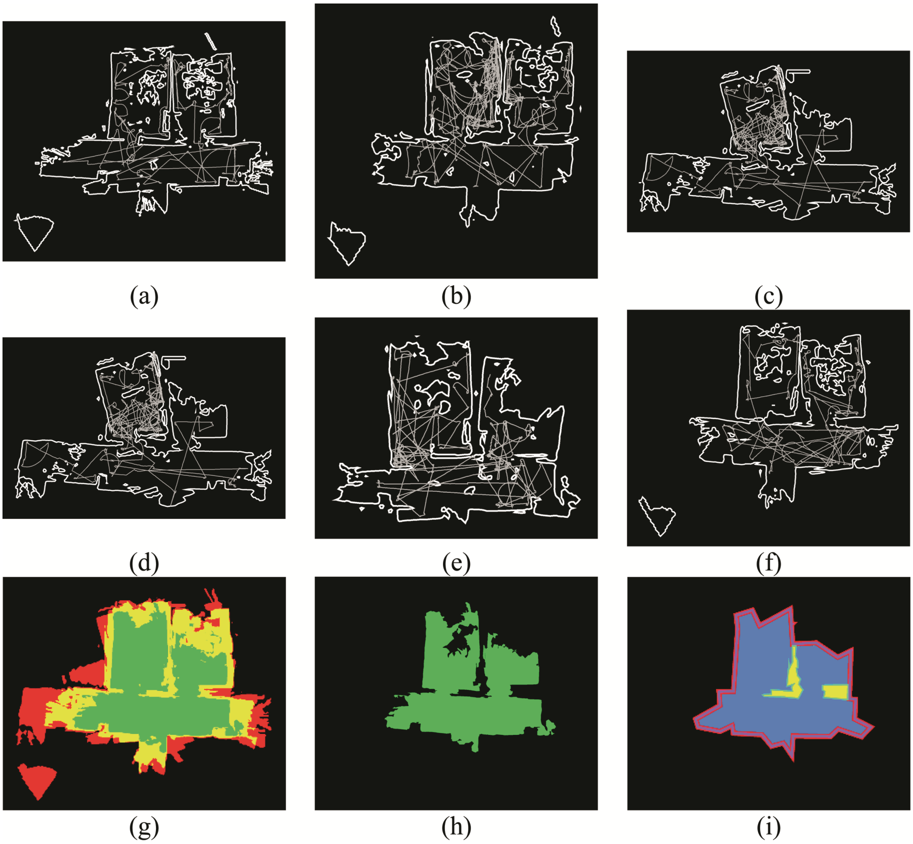
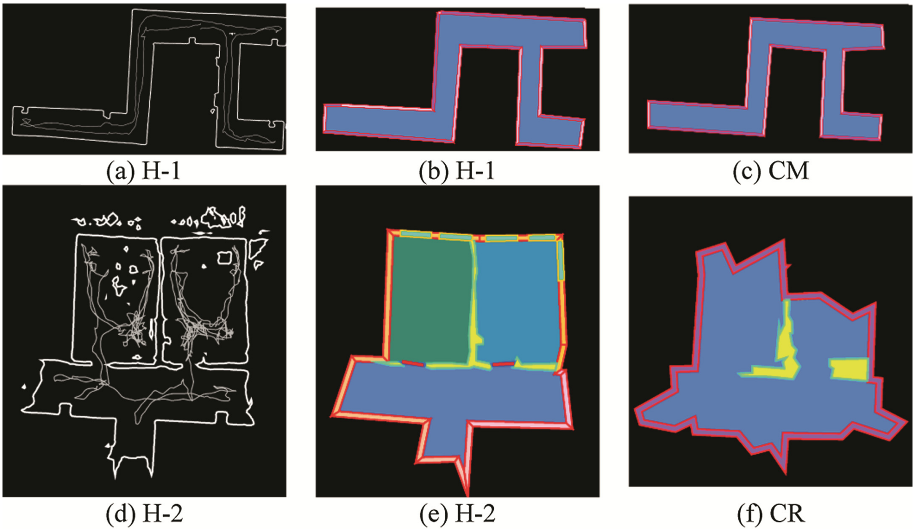


Fig. 14. Results of test scenario "CR", showing (a) – (f) the raw scan data of 6 individual scans, (g) the coverage heat map, (h) the cleaned polygon and (i) the reconstructed floor plan.

Figure 15 compares the results of the IoT-scenarios to human-made scans. We can see that the reconstruction result based on the usage of a cleaning machine (scenario "CM") is very close to the one achieved by a human. However, quite obvious differences can be observed when comparing the result of scenario "CR" (cleaning robot) to the human-made scan and reconstruction. Based on the scan conducted by human it was possible for the reconstruction pipeline to correctly capture the spatial topology of the scene and to detect all three rooms, the two doors connecting the offices with the hallway and also five windows.



(a) H-1 (b) H-1 (c) CM

(d) H-2 (e) H-2 (f) CR

Fig. 15. Comparison of IoT-scenarios "CM" and "CR" with human-made scans, showing on the left the raw scan data of the human-made scans, in the middle the reconstruction results based on these scans and on the right the corresponding results of the scenarios "CM" (c) and "CR" (f).

6 Conclusion and Outlook

In this paper we proposed a framework for autonomous BIM reconstruction using an IoT-based approach. We investigated the utilization of cleaning devices as drones for an indoor scanning process. Our hypothesis was that by using the recurrent movement of these devices within a building it would be possible to obtain multiple datasets without time consuming human interaction. We further investigated how the analysis of multiple scans can enhance the reconstruction result.

Our study showed that it is generally possible to use common devices like cleaning machines and cleaning robots as drones for indoor scanning. By doing so it is possible to drastically reduce the human workload. The coverage analysis of multiple scans as demonstrated in this study proved to be a suitable tool to make use of recurring scans and to enhance the data quality. However, the use of drones and autonomous devices also has some obvious drawbacks. Especially the cleaning robot showed to be prone to drift errors and often was not able to capture enough data to obtain reasonable reconstruction results. On the other hand, in a simpler environment, scan data obtained by mounting the depth sensor onto a cleaning machine showed satisfying reconstruction results. It becomes clear that the investigated hardware setups lack intelligent behavior. Humans are able to recognize if parts of a scene have not been sufficiently scanned and can react accordingly. In our current hardware setup however, there is no feedback between the sensor and the drone. The cleaning robot did not 'know' that it was scanning the environment and its movements were random and not designed to accomplish this kind of task.

We therefore see a lot of potential for further research in this area. Future work should evaluate the use of other sensor technologies, some of which are mentioned in Sect. 3. It should further be investigated how these sensors can be integrated with the devices APIs to achieve a more intelligent behavior. Additionally, efforts should be made to enable these devices to share their information directly e.g., communicate to each other which parts of the building have not yet been scanned sufficiently.

References

1. Beetz, J., et al.: Enrichment and preservation of architectural knowledge. In: Münster, S., Pfarr-Harfst, M., Kuroczyński, P., Ioannides, M. (eds.) 3D Research Challenges in Cultural Heritage II. LNCS, vol. 10025, pp. 231–255. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47647-6_11
2. Franz, S., Rüppel, U.: Supporting forensic fire cause analysis with next generation mobile devices by assembling a BIM based on multiple integrated sensor data. In: Proceedings of the 22nd International Workshop on Intelligent Computing in Engineering (2015). http://eg-ice-2015.bwk.tue.nl/proceedings_online.html
3. Franz, S., Irmler, R., Rüppel, U.: Collaborative BIM reconstruction on mobile consumer devices. In: Koch, C. (ed.) Digital Proceedings of the 24th International Workshop of the European Group for Intelligent Computing in Engineering, pp. 116–125 (2016). http://www.proceedings.com/35076.html
4. Nordrum, A.: Popular Internet of Things Forecast of 50 Billion Devices by 2020 is Outdated. IEEE (2016)
5. Wu, B., Yu, B., Wu, Q., Yao, S., Zhao, F., Mao, W., Wu, J.: A graph-based approach for 3D building model reconstruction from airborne LiDAR point clouds. **9**, 92 (2017). https://doi.org/10.3390/rs9010092
6. Tuttas, S., Stilla, U.: Window detection in sparse point clouds using indoor points. In: ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVIII-3-W22, Copernicus GmbH, pp. 131–136 (2013). https://doi.org/10.5194/isprsarchives-XXXVIII-3-W22-131-2011, https://www.int-arch-photogramm-remote-sensspatial-inf-sci.net/XXXVIII-3-W22/131/2011/
7. Sohn, G., Huang, X., Tao, X.: Using a binary space partitioning tree for reconstructing polyhedral building models from airborne lidar data. Photogrammetric Engineering and Remote (2008)
8. Suveg, I., Vosselman, G.: Reconstruction of 3D building models from aerial images and maps **58**(3), 202–224 (2004). https://doi.org/10.1016/j.isprsjprs.2003.09.006, http://www.sciencedirect.com/science/article/pii/S0924271603000583
9. Previtali, M., Scaioni, M., Barazzetti, L., Brumana, R.: A flexible methodology for outdoor/indoor building reconstruction from occluded point clouds. In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-3, Copernicus GmbH, pp. 119–126 (2014). https://doi.org/10.5194/isprsannals-II-3-119-2014, https://www.isprs-ann-photogramm-remote-sens-spatial-nf-sci.net/II-3/119/2014/
10. Koch, T., Körner, M., Fraundorfer, F.: Automatic alignment of indoor and outdoor building models using 3D line segments. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 689–697. https://doi.org/10.1109/cvprw.2016.91
11. Strecha, C., Krull, M., Betschart, S.: The Chillon Project: Aerial/Terrestrial and Indoor Integration (2014). https://pix4d.com/wp-content/uploads/2016/03/Pix4D-White-Paper-Chillon-Project-June2014.pdf

12. Schindler, K., Bauer, J.: A model-based method for building reconstruction. In: First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, HLK 2003, pp. 74–82 (2003). https://doi.org/10.1109/hlk.2003.1240861
13. Ochmann, S., Vock, R., Wessel, R., Klein, R.: Automatic reconstruction of parametric building models from indoor point clouds. **54**, 94–103 (2016). https://doi.org/10.1016/j.cag.2015.07.008, http://www.sciencedirect.com/science/article/pii/S0097849315001119
14. Turner, E., Cheng, P., Zakhor, A.: Fast, automated, scalable generation of textured 3D models of indoor. Environments **9**(3), 409–421 (2015). https://doi.org/10.1109/JSTSP.2014.2381153
15. Corso, N., Zakhor, A.: Indoor localization algorithms for an ambulatory human operated 3D mobile mapping system. 5 (12), 6611–6646 (2013). https://doi.org/10.3390/rs5126611, http://www.mdpi.com/2072-4292/5/12/6611
16. Dumitru, R.-C., Borrmann, D., Nüchter, A.: Interior reconstruction using the 3D hough transform. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-5-W1, Copernicus GmbH, pp. 65–72 (2013). https://doi.org/10.5194/isprsarchives-XL-5-W1-65-2013, https://www.int-arch-photogramm-remote-sensspatial-inf-sci.net/XL-5-W1/65/2013/
17. Du, H., Henry, P., Ren, X., Cheng, M., Goldman, D., Seitz, S.M., Fox, D.: Interactive 3D modeling of indoor environments with a consumer depth camera. In: UbiComp 2011 - Proceedings of the 2011 ACM Conference on Ubiquitous Computing, pp. 75–84 (2011). https://doi.org/10.1145/2030112.2030123
18. Yue, H., Chen, W., Wu, X., Liu, J.: Fast 3D modeling in complex environments using a single Kinect sensor. **53**, 104–111 (2014). https://doi.org/10.1016/j.optlaseng.2013.08.009, https://www.sciencedirect.com/science/article/pii/S0143816613002522
19. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGB-D mapping: using Kinect-style depth cameras for dense 3D modeling of indoor environments **31**(5), 647–663 (2012). https://doi.org/10.1177/0278364911434148
20. Miljanovic, M., Eiter, T., Egly, U.: Detection of windows in facades using image processing algorithms (2012)
21. Xiong, X., Adan, A., Akinci, B., Huber, D.: Automatic creation of semantically rich 3D building models from laser scanner data. In: Automation in Construction, vol. 31 (2013). https://doi.org/10.1016/j.autcon.2012.10.006
22. Okorn, B., Xiong, X., Akinci, B., Huber, D.: Toward automated modeling of floor plans (2010)
23. Cabral, R., Furukawa, Y.: Piecewise planar and compact floorplan reconstruction from images. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 628–635. IEEE Computer Society (2014). https://doi.org/10.1109/cvpr.2014.546
24. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Reconstructing building interiors from images. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 80–87 (2009). https://doi.org/10.1109/iccv.2009.5459145
25. Zhou, Q.-Y., Koltun, V.: Color map optimization for 3D reconstruction with consumer depth cameras. **33** (4), 155:1–155:10 (2014). https://doi.org/10.1145/2601097.2601134
26. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality, pp. 127–136 (2011). https://doi.org/10.1109/ismar.2011.6092378
27. Bormann, R., Jordan, F., Li, W., Hampp, J., Hägele, M.: Room segmentation: survey, implementation, and analysis. In: 2016 IEEE International Conference on Robotics and Automation (ICRA) (2016)

28. Kleiner, A., Rodrigo Baravalle, R., Kolling, A., Pilotti, P., Munich, M.: A solution to room-by-room coverage for autonomous cleaning robots. In: Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada (2017)
29. Ackerman, E.: Neato Adds Persistent, Actionable Maps to New D7 Robot Vacuum. https://spectrum.ieee.org/automaton/robotics/home-robots/neato-adds-persistent-actionable-maps-to-new-d7-robot-vacuum. Accessed 20 Dec 2017
30. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors **12**(2), 1437–1454 (2012). https://doi.org/10.3390/s120201437
31. Chang, C.-Y., Langel, B., Zhang, M., Koenig, S., Requejo, P., Somboon, N., Sawchuk, A.A., Rizzol, A.A.: Towards pervasive physical rehabilitation using Microsoft Kinect. In: Proceedings of the 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops (2012)
32. Gabel, M., Gilad-Bachrach, R., Renshaw, E., Schuster, A.: Full body gait analysis with Kinect. In: Proceedings of the 2012 Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC) (2012). https://doi.org/10.1109/embc.2012.6346340
33. Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., Presti, P.: American sign language recognition with the kinect. In: 2011 Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI) (2012). https://doi.org/10.1145/2070481.2070532
34. Stowers, M., Hayes, M., Bainbridge-Smith, A.: Altitude control of a quadrotor helicopter using depth map from Microsoft Kinect sensor. In: Proceedings of the 2011 IEEE International Conference on Mechatronics (ICM) (2012). https://doi.org/10.1109/icmech.2011.5971311
35. Product Brief REAL3 image sensor family|infineon. 11|2015. https://www.infineon.com/dgdl/Infineon-REAL3\%20Image\%20Sensor\%20Family-PB-v01_00-EN.PDF?fileId=5546d462518ffd850151a0afc2302a58. Accessed 20 Dec 2017
36. Asus ZenFone AR (ZS571KL)|Asus. https://www.asus.com/Phone/ZenFone-AR-ZS571KL/Tech-Specs/. Accessed 20 Dec 2017
37. Structure Sensor and Structure SDK Developer Information. https://structure.io/developers, Accessed 20 Dec 2017
38. Orbbec Persee. https://orbbec3d.com/product-persee/. Accessed 20 Dec 2017
39. Development Kit Brief CamBoard pico flexx|pmd. http://pmdtec.com/picofamily/assets/datasheet/Data-sheet-PMD_RD_Brief_CB_pico_flexx_V0201.pdf. Accessed 20 Dec 2017
40. Intel RealSense Depth Camera D435. https://click.intel.com/intelr-realsensetm-depth-camera-d435.html Accessed 20 Dec 2017
41. ZED Mini|Stereolabs. https://www.stereolabs.com/zed-mini/. Accessed 20 Dec 2017
42. Visvalingam, M., Whyatt, J.D.: Line generalisation by repeated elimination of points. **30**(1), 46–51. https://doi.org/10.1179/000870493786962263 (1993)

Space Classification from Point Clouds of Indoor Environments Based on Reconstructed Topology

Wolfgang Huhnt[1](✉), Timo Hartmann[1], and Georg Suter[2]

[1] Technische Universität Berlin, Berlin, Germany
{wolfgang.huhnt, timo.hartmann}@tu-berlin.de
[2] Technische Universität Wien, Wien, Austria
georg.suter@tuwien.ac.at

Abstract. Reconstruction of as-built building models from point cloud data is a challenging problem with promising applications in the construction industry. In this paper, we outline the general concept of a data processing pipeline that produces fully three-dimensional, semantically rich and topologically valid as-built building models. Point cloud data is processed with a combination of histogram, voxel-based and RANSAC-based methods to detect surfaces of spaces and building components. Topological relations between building components (walls, slabs) are derived from a space partitioning that is generated from detected surfaces. The output from topology reconstruction is used as input for a space classification procedure which involves assigning functional properties to spaces. Each step in the data processing pipeline is illustrated with examples. Limitations of the proposed approach are discussed and an outlook of future development in this area is given.

Keywords: Building model reconstruction · Geometric modeling
Building information modeling

1 Introduction

The European building stock is deteriorating at a fast rate. Most of the existing buildings do no longer satisfy current user needs, do not match newer regulations, and are highly energy inefficient. Because of the high density of our urban environments and the large amount of embodied energy within the existing building stock, it will be important to upgrade existing buildings instead of demolishing them. To strategically decide which buildings within an overall building asset should be renovated, but also to design specific renovation measures requires a good understanding of existing conditions of buildings. To this end, accurate as-built descriptions of buildings are required. In this context, the term "as-built" refers to capturing sufficiently detailed depictions of the state of a facility as it was actually built or as it exists currently. As-built models are considered as a much better ground truth than the as-designed models frequently used in practice that are based on outdated construction drawings.

It is not surprising that the last years have seen an increased academic interest in as-built reconstruction research based on point clouds generated from advanced

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 82–102, 2018.
https://doi.org/10.1007/978-3-319-91635-4_5

photogrammetric sensor systems. The main focus of these research approaches has been on the reconstruction of the geometric features of buildings. Recently insights have also been developed for the automated recognition of objects from reconstructed geometry, such as walls, windows, ceilings, and doors. One aspect that has, to the best of our knowledge, not been addressed adequately yet is the reconstruction of a building's topology. Topology, however, is important for many automated assessment tasks related to understand the renovation requirements for buildings.

Objects of and in a building can be subdivided into four categories:

- building components (such as walls and ceilings),
- built-in components (such as doors and windows),
- furnishing and circulation elements (such as kitchen equipment and stairs), and
- spaces (such as rooms or hallways).

Furthermore, for many analysis applications the exterior space also needs to be considered.

Building components and spaces including their neighboring relations among each other are important. This includes contact surfaces between building components which are not visible from outside. They cannot be detected by visual inspections and laser scans; but they are important for various applications.

Present research in the field of automated as-built model reconstruction addresses the detection of building components and spaces (Mura et al. 2014). Additionally, relations between built-in components and building components, as well as, relations of furnishing and circulation elements to spaces are addressed in existing approaches. However, to the best of our knowledge, no approaches exist yet which addresses relations between building components and spaces in the 3-dimensional space including neighboring relations at invisible contact surfaces of building components.

Information about space functions related to user activities is required or desirable for many applications of as-built building models. Examples include architectural design, remodeling, and room space measurement (GSA 2009). However, existing work in as-built building modeling is limited to the reconstruction of space geometry and connectivity relations between spaces. Apparently, automated derivation of functional space properties, or space classification, so far has not been attempted. We select the domain of multi-unit residential buildings in this research.

To overcome these gaps, we outline a method for creating as-built topological models of existing buildings from point cloud data of building interiors and the exterior. We propose a data processing pipeline that consists of three major steps (Fig. 1). First, registered point clouds are processed to produce surfaces of building components and built-in components. Furnishing and circulation elements are at present time not considered. Second, a topologically correct representation is created that models the objects' geometries and how objects are connected to each other. Third, this topology representation is used to classify spaces according to their functional properties.

The pipeline starts with processing a registered point cloud to identify the characteristic geometries and objects of a building. To this end, we suggest to use a combination of existing point cloud processing techniques. Our method builds upon histogram based approaches combined with line tracing algorithms to identify a

Point cloud processing
Topology reconstruction
Space classification

Fig. 1. Steps in space classification from point clouds.

building's shell and room partitioning surfaces. Finally, we propose to use the initially identified surfaces as seeds for stochastic RANSAC type algorithms to distinguish between building components, built-in components, and to increase the accuracy of the identified surfaces.

As a result, the initial point cloud methods can provide a basic geometrical description of the building that can then be used to identify the building's topology.

The next step is to compute a correct topology of building components and rooms. Based on given surfaces, objects including their contact surfaces are reconstructed. The proposed approach is based on space partitioning. The concept of twins is used for surfaces and edges.

As a result, the topology is explicitly available as integral part of the geometry. Based on this information, graphs are derived that describe the topology between building components and rooms. These graphs including geometrical information such as areas and volumes form the input for space classification.

In the last step of the pipeline, spaces are classified. That is, functional properties of spaces as well as the functional units to which they belong are derived. Together with floor area and space volume properties, these data are useful e.g. for decision making in building renovations.

This paper presents the general concept. We start with an overview of the general concept in Sect. 2. Then, each major step is presented in the following three sections. The paper ends with conclusions and an outlook.

2 General Concept

Each step in the data processing pipeline requires input data and produces new data which is used as input to subsequent steps. An overview of the data flow in the data processing pipeline is given in Fig. 2.

The data processing pipeline accepts as input a registered point cloud of a building's indoor and outdoor environment. The point cloud processing step produces collections of surfaces of building components such as walls, ceilings, etc. Doors and windows are detected and associated with surfaces in which they are contained. In addition, two different kinds of relations of surfaces are detected: neighboring relations of two surfaces which describe the existence of a common edge, and neighboring relations of two surfaces which describe that they are surfaces of the same building component such as at the front and at the rear or the top and the bottom.

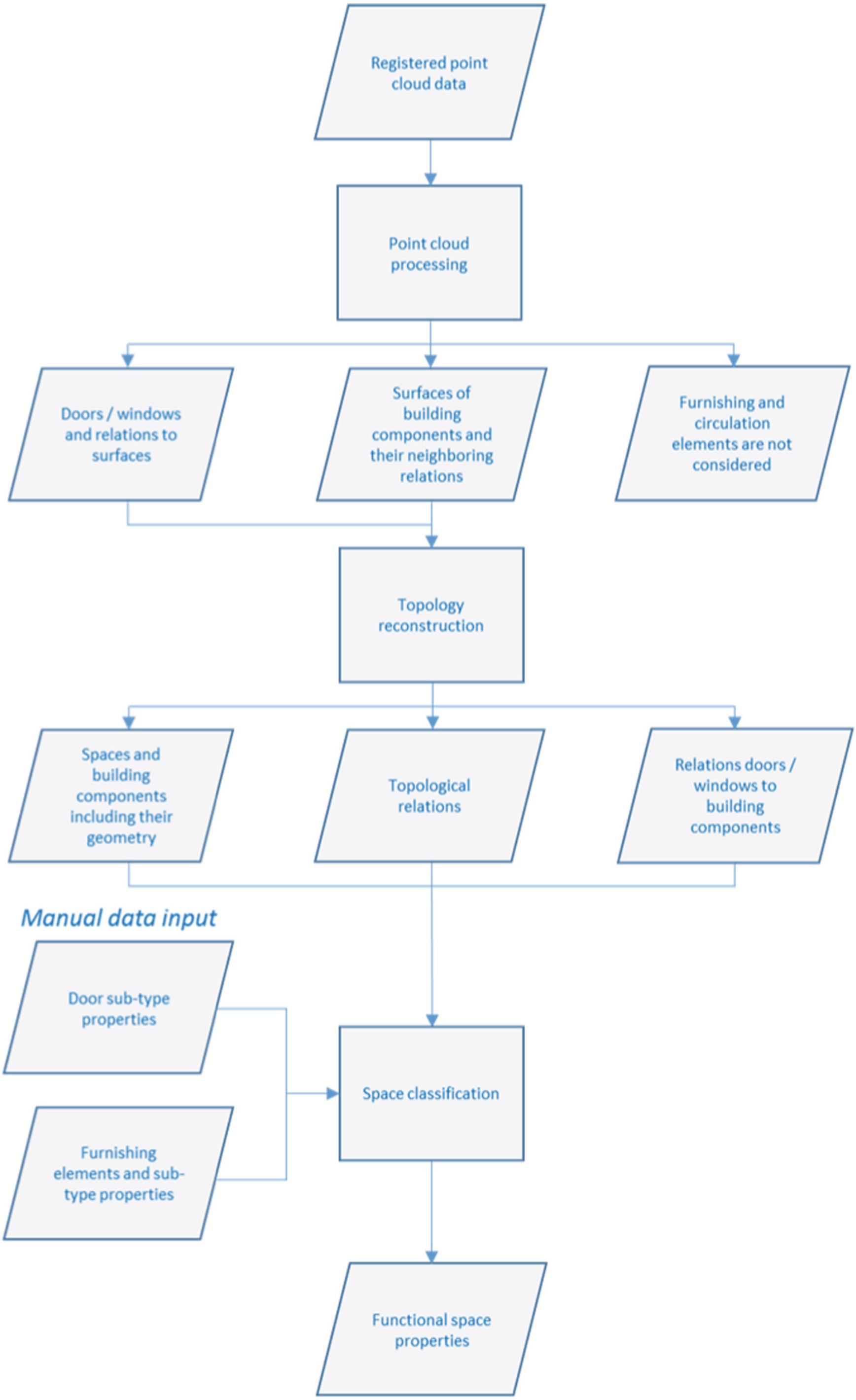

Registered point cloud data
Point cloud processing
Doors / windows and relations to surfaces
Surfaces of building components and their neighboring relations
Furnishing and circulation elements are not considered
Topology reconstruction
Spaces and building components including their geometry
Topological relations
Relations doors / windows to building components
Manual data input
Door sub-type properties
Furnishing elements and sub-type properties
Space classification
Functional space properties

Fig. 2. Data flow in the data processing pipeline.

Topology reconstruction identifies building components and spaces. In addition, their geometry is determined; and relations between building components and spaces are identified. The geometry of spaces is used to calculate floor areas. Relations of doors and windows to building components result from their relations to the surfaces of these components.

Space classification accepts space data as input from topology reconstruction. Door, window, furnishing and circulation element data are required together with connectivity relations between these objects and spaces. Missing semantic data for doors, furnishing and circulation elements are assumed to be added manually.

3 Point Cloud Processing

3.1 Overview

Currently, the two major methods to identify surfaces from point cloud scans are based upon histograms or the RANSAC algorithm. Essentially, both methods are of statistical nature and are steered by additional parameters that need to be adjusted to the situation at hand. The next subsections describe the two approaches in more detail.

3.2 Histogram and Voxel Based Approaches

Histogram based methods work by projecting laser scan points on a surface and then counting the number of points that fall into a specific patch on this surface. The counts can then be used to understand the height of the scanned objects that are perpendicular to the projection surface. For patches with a large point count one can understand whether surfaces are located above this patch and the relative length of the surface. In the context of buildings where the additional assumption can be made that walls are perpendicular objects that originate from floor surfaces and that reach over the entire height of a room histogram based methods can be used to identify the bounding surfaces of a laser scanned building (Jung et al. 2014; Sietzema 2015).

To this end, a first segmentation of the point cloud can be conducted by projecting points to one of the vertical planes. The patches of the resulting histogram with the largest number of points can then be considered as patches that are on the height of the ceiling of a building. Of course, this is only valid under the assumption that ceilings stretch the entire length of a building or, in other words, that the building does not contain any large vertical spaces that stretch over multiple levels, such as atriums. For such buildings, histogram-based methods can also be used, however, here a threshold needs to be defined that specifies upon which number of points per patch a level boundary is assumed, that only stretches over a part of the entire length of the building.

In both cases, using the information about the location of the ceiling, the overall point cloud can be segmented into several point clouds representing each level. For each of the point clouds, in turn, a new histogram can be generated, this time projecting the points on the horizontal surface. Following the same logic as for the identification of ceilings, those patches with a maximum point count can be considered as patches from which walls originate. Similar to the identification of levels, problems occur with furnishing and circulation elements that do not cover the entire height of a space, such as counters or other furniture elements. Other than for the detection of ceilings these elements can usually be filtered well. Likewise, the amount of points that describe the elements are much fewer than the amount of those patches that describe walls (Fig. 3).

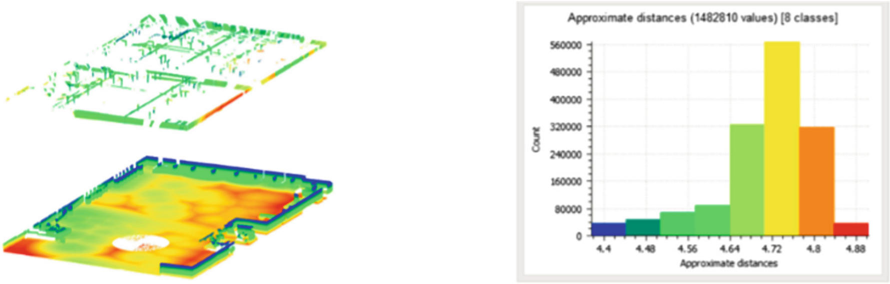


Fig. 3. Example of a histogram based identification of walls. Based on a count of the projected points on the horizontal surface, wall surfaces can be identified.

The identified patches within the histograms, can then be combined to recreate the main surfaces of the building. For example, walls can be reconstructed from the patches within the horizontal surface together with the information about the height of the building from the identification of the levels.

Voxel based method transform the concepts of the histogram based methods to three dimensions (3D). These methods subdivide the entire 3D point cloud space in equal cubes, so called voxels. The number of points contained in each of the voxels is then counted and used to decide whether a voxel represents a solid object or not. This information can then also be used to reconstruct the bounding surfaces of a building, often using machine learning based approaches (Xiong et al. 2013; Wang and Tseng 2011).

3.3 RANSAC Based Approaches

The random sample consensus (RANSAC) algorithm is an iterative method to reconstruct a surface from an initially selected small set of points. These points are usually randomly selected. The algorithm then checks which of the points are logically related to this initial set of points using characteristics of the initial point patch, such as location or normal vector.

For example, assuming that wall or ceiling surfaces are objects without a curvature, the RANSAC algorithm can identify those points that are close to the initial patch and that would be within the normal plane of the initially chosen points. To this end, a threshold needs to be defined for both the distance of points from the initial patch that should be included and for the maximum angle that this point would deviate from the normal of the initial points.

RANSAC based approaches are powerful methods in cases an initial set of points can be clearly identified for each surface to reconstruct. Therefore, RANSAC approaches do not lend themselves very well for the identification of surfaces from cluttered initial point clouds. They are much better suited to clearly delineate and reconstruct surfaces for which already a number of initial points are known (Wang et al. 2015).

Because of this reason RANSAC algorithms are often only used as a second step after an initial point cloud has already been dissected using the above described histogram and voxel based methods. The next section describes such potential for combining different methods and approaches.

3.4 Combined Methods and Challenges

To clearly delineate the surfaces of a building, a combination of the two above described major point cloud processing methods yield quite some potential. Histogram based methods allow to clearly reconstruct the basic skeleton of a building. In this way, point clouds of whole buildings can be dissected in parts, such as the above-mentioned building levels. Additionally, these methods help to remove clutter from the point clouds, or in other words points that do not represent surfaces of the building's skeleton, such as surfaces delineating furniture or other equipment. However, the geometrical accuracy of these efforts often lack precision. Therefore, histogram based methods can be meaningfully combined with RANSAC approaches to increase the geometrical accuracy of the reconstruction for wall and ceiling surfaces. RANSAC algorithm can be started based on point patches that are representative for the initially identified surfaces, significantly increasing the accuracy of the reconstruction effort.

Next to these approaches, RANSAC algorithms also allow well for the identification of openings within these surfaces, such as doors and windows. Based on an appropriate selection of a seed patch, RANSAC algorithms often allow for a very clear delineation between surfaces that represent walls and surfaces that represent window or door panes (Yousefzadeh 2016).

Next to the above combination of methods, additional information from the scan can be used to further increase the accuracy of the selected surfaces. For example, to scan the indoors of an entire building usually one scan needs to be done from within every room of the building. These different scans are only later locally referenced with each into a combined point cloud, however they can be compared initially with each other. For example, parallel histogram lines in different scans from different scans that are in close proximity within the combined point cloud can be assumed to represent the surfaces of a single wall.

While accurate software to reconstruct the surfaces that represent building components can be implemented using the above summarized methods, one problem that by and large still exists is the reconstruction of the complete topological configuration of a building. The next section will describe a method for this reconstruction.

4 Topology Reconstruction

4.1 Introduction

Different topological relationships can exist between objects, and these relationships can be mathematically described (Egenhofer Herring 1990). Research has been con-ducted in this field. Nguyen et al. developed algorithms based on AutoCAD to deduced topological relations from a given boundary representation (Nguyen et al. 2005). Their work is restricted to simple geometries. Borrmann and Rank present a query language with the goal to extract topological relations from a given boundary representation (Borrman and Rank 2009(a)). In (Borrman and Rank 2009(b)), they present topological operators. Their work is based on an octree data structure. The authors expand the query language in (Borrman et al. 2009(c)) to the computation of distances between objects.

In the context of the research presented in this paper, the geometry itself is not completely given as it is reconstructed from point clouds. Therefore, the extraction of topological relations requires an up-stream activity: a decomposition into objects. In this context, building components and spaces are important. Based on such a model, an algorithm must extract relevant topological relations between building components and spaces.

Challenges are the identification of building components and the correct geometrical determination of contact faces. The topological relationship is explicit once contact faces are correct and modeled in such a way that they know the two neighboring building components. The inherent problem is a question of the actual point of view. Figure 4 shows in (1) visible surfaces of two walls which can be the corner of a room. The descriptions of these two walls can be done differently due to the fact that the corner might belong to one wall, (2) and (3), or might be a separate wall (4). All resulting three different models differ in geometry and topology.

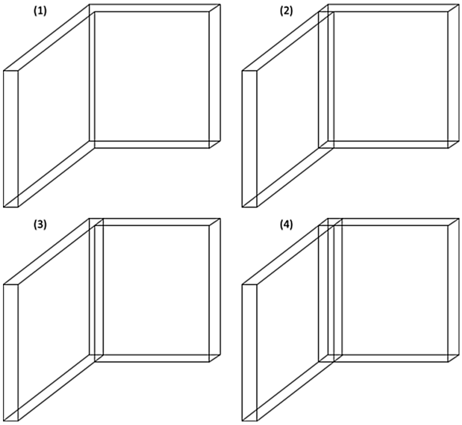


Fig. 4. Two walls.

The approach presented in this paper is based on partitioning the space. The idea of a partitioning of space is to model n−1-dimensional objects which separate n-dimensional objects. In (Mäntylä 1988), a specific data structure Half Edge (HE) is introduced to model polygonally bounded objects. In contrast to the former work of (Baumgart 1972), edges are split by their direction. Two half edges are representatives of a single undirected edge. These edges cover information about both neighboring objects. They partition a 2D space.

The concept of half edges can be adapted to faces which are oriented in 3D space. Two half faces can be introduced to separate two three-dimensional objects. Each half face can refer to one three-dimensional object; and the two half faces of two neighboring three-dimensional objects know each other for navigation purpose. The term twin is often used to describe the reference of half objects to each other.

In this research, faces which have been generated from captured point clouds form the input. These faces separate building components from interior spaces or the exterior. In dimensional reduced approaches, faces can represent for instance two sides of a wall. Here we are dealing with 3-dimensional building components. Faces from point clouds are used to generate both, surfaces of building components and surfaces of interior space and the exterior. In addition, contact faces between building components are generated. During this process, the algorithms determine the detailed description. Thus, returning to Fig. 4, the algorithms identify three building components with two invisible contact faces (4). The resulting structure is a partitioning of space which covers building components, interior spaces and the exterior. This structure is analyzed in a second step. The concept is illustrated in the next sub-section.

4.2 Concept

This section describes the concept for topology reconstruction. The concept itself is structured into four steps. It is based on a preprocessing step of point cloud processing.

Preprocessing Step. Required results of point cloud processing are illustrated in Fig. 5. Figure 5 shows a part of a building where two walls meet. This example is chosen in this subsection to illustrate the concept of topology reconstruction. The example consists of two objects; but the principle is transferrable to any number of objects.

Point clouds form the basis for calculating surfaces of objects (Fig. 5(1) and (2)) as described in Sect. 3. Two different kinds of relations are necessary. One kind describes that surfaces touch each other at a common edge, the other kind describes that surfaces belong to the same object (Fig. 5(3) and (4)). The first kind of relationship results directly from point cloud processing as described in Sect. 3. The second kind of relationship requires a calculation where the distance between faces and their normal are used to identify that different faces are surfaces of the same building component.

Step 1: Generating Input Geometry for Space Partitioning. Space partitioning as used in this approach requires an input geometry for each building component. The approach presented in this paper addresses not only visible surfaces of building components; it also addresses invisible contact faces. For this purpose, surfaces generated from point clouds are extended. Additional surfaces are added so that the reconstruction

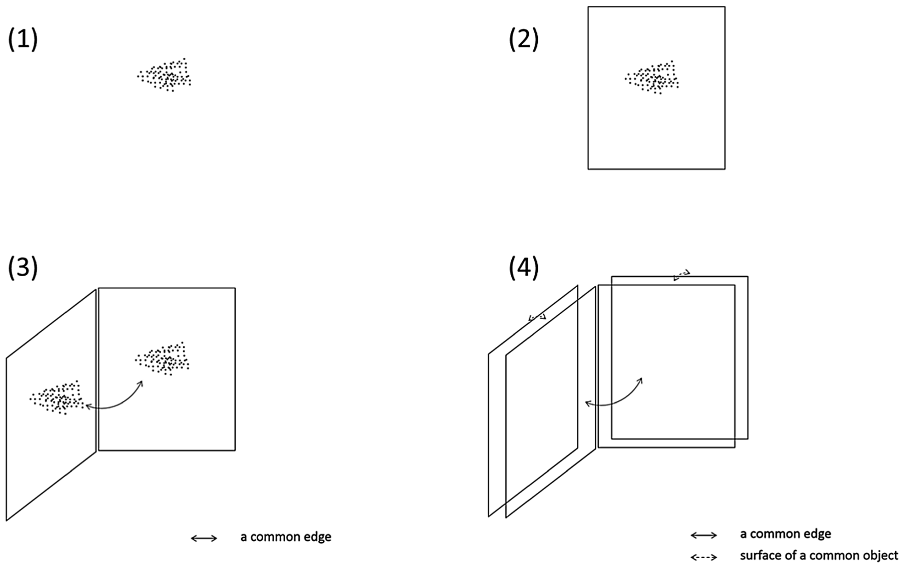


Fig. 5. From point clouds to surfaces and relations between surfaces.

starts with a valid solid for each building component. The geometry of this solid is at that point in time not the correct geometry of the building component. In addition, contact faces between building components are not known at that point in time

Both kinds of relations between faces are evaluated for the generation of an input geometry. The size is chosen in such a way that the input geometry for a building component is always sufficiently large such that it extends beyond its neighboring objects. This is illustrated in Fig. 6. Two building components meet at a corner, and the surfaces of both components are extended in such a way that they protrude over the corner.

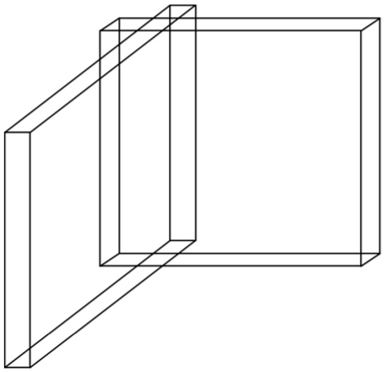


Fig. 6. Input for space partitioning.

This concept is applied to each convex corner where different surfaces meet. The rear side need not be considered due to the fact that it results from the lengthening. Rear sides typically have a concave geometry if surfaces of building components are parallel.

Step 2: Space Partitioning. In this step, specific data structures are generated. These data structures are based on the consideration that each 2-dimensional face object knows the two 3-dimensional objects that are separated by this face. The algorithm is done in two major steps. In a first step, all edges of given objects are reconstructed. This includes the determination of intersection points of given edges as shown on the upper and the lower faces in Fig. 7 (Huhnt 2018). The second mayor step is the reconstruction of faces and volumes. This includes the determination of intersections of different given surfaces. Figure 7 shows four additional edges compared to Fig. 6. These edges are intersection edges of the two solids. These edges are cutting edges because the two given objects penetrate each other. These edges are determined during this step.

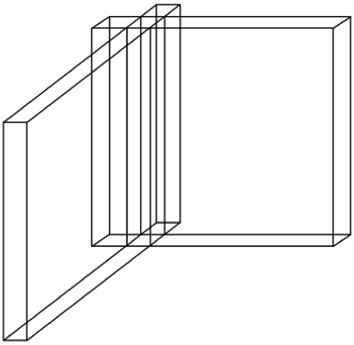


Fig. 7. Result of space partitioning.

Step 3: Analyzing the Resulting Structure. As shown in Fig. 8, different volumes are allocated to different objects. The exterior and one interior are added to illustrate that a corner of two walls is only a part of a building. In reality, several volumes are generated in the interior of a building which forms the different rooms. The resulting structure is analyzed concerning the allocation to objects. First of all, space can be allocated by given building components or voids. Voids can be interior space or the exterior. Voids are identified by the fact that they are not allocated to any building component. Voids are separated by the fact that they are not connected. The analysis concerning the allocation to given building components results in two main categories:

1 space which is allocated to a single building component, and
2 space which is allocated to more than one building component.

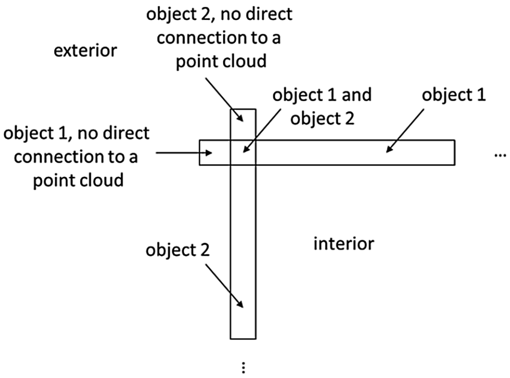

exterior
object 2, no direct connection to a point cloud
object 1 and object 2
object 1
object 1, no direct connection to a point cloud
...
object 2
interior
⋮

Fig. 8. Analyzing results of space partitioning

Space which is allocated to a single building component is split into two subcategories:

1a space which is connected to a point cloud, and
1b space which is not connected to a point cloud.

Figure 8 shows a result of such an analysis. For further investigations, space of the category 1b is not of interest. This space is allocated to its neighboring void. It resulted from extending of captured surfaces as described in step 1. The resulting structure describes the geometry of all given objects including contact faces between objects. An example is shown in Fig. 9.

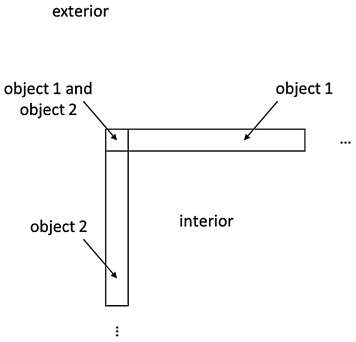


Fig. 9. Reconstructed geometry.

Step 4: Extracting Topological Relations. The last step is the extraction of topological relations. These relations are determined by the evaluation of neighboring relations of space partitioning and allocation to building components and spaces. An example is shown in Fig. 10.

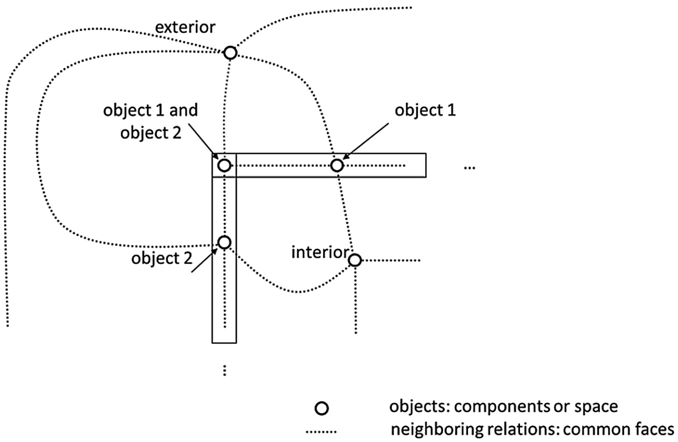

exterior
object 1 and object 2
object 1
object 2
interior
...
○ objects: components or space
.... neighboring relations: common faces

Fig. 10. Topological relations.

4.3 Algorithms

Kinds of Algorithms. The reconstruction requires three different kinds of algorithms. These three different kinds of algorithms are shortly summarized in the following subsections.

The core algorithm is the reconstruction of all objects based on space partitioning. In (Kraft 2016), a first approach was presented which was able to handle several objects. The approach developed by Kraft uses floating point values for coordinates. The disadvantage of this approach is that a refinement is necessary. Another approach based on integer values are coordinates is in progress (Huhnt 2018) with the goal to avoid refinements.

Generating Input Geometry. The generation of input geometry requires data from point cloud processing. Surfaces including their normal are direct results of point cloud processing. Neighboring relations between surfaces are also results of point cloud processing.

The generation of an input geometry for building components uses at its first step distances between point clouds and the normal of each surface to decide whether surfaces belong to the identical object. An octree data structure is the proposed solution strategy. The application of octrees to digital building models has already be investigated in (Borrman and Rank 2009 (1)) for the calculation of topological relationships. Identified candidates are checked whether the normal of the two surfaces do not point in the same direction.

In a second step, valid solids are generated for each building component. At present time, we use a watertight and not self-intersecting set of triangles as objects to describe the boundary of each building component. The normal of each triangle defines the exterior of the associated building component.

Reconstruction. The reconstruction requires a valid space partitioning as input. This is achieved by a mesh of tetrahedrons where vertices have integer values as coordinates. Neighboring tetrahedrons are stored for navigation purposes. Integer values guarantee reproducible geometric test.

In a first step, all vertices are inserted into an initial mesh of tetrahedrons. Split-operators guarantee that the mesh is always composed from tetrahedrons. In a second step, all edges of all given triangles are reconstructed. Intersection points between the tetrahedral mesh and these edges are computed.

The following step is the reconstruction of the triangles. For this purpose, each affected tetrahedron is decomposed into a set of polyhedrons. A surface of a polyhedron is described by a polygon. Each polygon knows the building component which is located underneath. This information is transferred to the polyhedron which is underneath that polygon.

The resulting structure is analyzed. Sets of polyhedrons form building components and volumes which belong to more than one building component. Spaces in the interior of a building are volumes which are not allocated to any building component. Spaces in the interior do not have any connection to the exterior. Spaces in the interior are not interconnected volumes.

Analysis of Space Partitioning The most important algorithms for the analysis of the set of polyhedrons is the Breadth-First-Search. The Breadth-First-Search is a well-known algorithm from graph theory. It is applied to the set of polyhedrons to identify the boundaries of building components and spaces.

Results for Space Classification. Space classification requires different sets which are results of topological reconstruction:

- a set of building components,
- a set of interior spaces, and
- the exterior.

Each element in this set knows its geometry so that for instance floor areas are available.

In addition, space classification requires the set of built-in components, doors and windows. This set is generated by point could processing.

Also relations are input for space classification. Relations in the and between the sets of building components and spaces are results of topology reconstruction. Relations of building components to built-in components result from point cloud processing. Point cloud processing identifies built-in components and their relationship to surfaces of building components. This relationship is used to assign built-in components to building components.

5 Space Classification

5.1 Overview

The richly structured semantic and geometry data produced by topology reconstruction is used for automated space classification. In the following, a procedure is described for classification of spaces in multi-unit residential buildings. This is a promising domain because of the prevalence of such buildings in urban environments and the portfolios of professional real estate owners and managers.

Existing space classification systems are based on main space functions that are related to user activities (OSCRE, CSI 2010). They are used in space management or floor area calculation (ANSI/BOMA 2010a, DIN 2016). The OmniClass system relies on a four level generalization hierarchy. For example, a ‘bedroom’ (level 2) is a subtype of ‘private residential space’ (level 1). The system includes 960 space types. However, coverage of residential spaces is presently limited.

Detailed, highly accurate space area calculations require significant domain knowledge and expertise. This is a manual, time intensive task which is difficult to automate and formalize, not least due to different calculation methods. By contrast, our aim is to facilitate automated area calculations that are sufficiently accurate to support decision-making for residential building portfolio management.

A subset (or view) is retrieved from the data generated by topology reconstruction. Space classification itself is done in two steps. First, a graph that models connectivity relations between windows, doors, and spaces is processed to extract residential and primary circulation units. In addition to space classification, the derivation of units enables floor area calculations at multiple aggregation levels. Second, functional properties are determined for spaces in each unit. Heuristics are based on semantic and geometric space properties.

5.2 Functional Units and Functional Space Properties

Calculation of floor areas in multi-unit residential buildings is done on at multiple aggregation levels (ANSI/BOMA 2010b, DIN 2004). Floor areas are computed at the level of individual spaces and functional units. In this work, a functional unit (FU) refers to a group of connected, mutually accessible spaces that serve specific functions. Two FU types are of particular interest in multi-unit residential buildings. Residential FUs (that is, apartments) provide private spaces for people living in households. Primary circulation FUs are shared spaces that provide access to residential units. Examples for spaces in primary circulation FUs are stairs, elevators, hallways, or building entrances. Multi-unit residential buildings typically include units for technical infrastructure, such as heating rooms, or units with shared spaces for communal or commercial activities and storage. We currently do not consider the derivation of such FUs and the classification of their spaces.

We use the concept of functional property (FP) to refer to space use. In other words, an FP represents a specific activity in a space. In order to model multi-functional spaces, we assume that a space has a primary FP and one or more secondary FPs. For example, the primary FP of a living room is to enable social activities. If a bedroom is accessed from it, then a secondary FP is to serve as a circulation space.

Detailed floor area analysis is feasible based on FPs of individual spaces or groups of spaces with similar or equivalent generalized FPs. In floor area calculations of residential units, usable floor area is distinguished from circulation and technical infrastructure floor areas (DIN 2004). More specific measures, such as floor areas of living rooms or bedrooms in residential FUs, may be useful for comparative analysis and decision making in the remodeling of residential buildings.

5.3 Space Classification View

Required data for space classification and their sources are listed in Table 1. Most data are used directly from topology reconstruction. For the derivation of FUs (Sect. 5.4), normal doors, unit doors, and elevator doors are distinguished. Normal doors connect spaces in an FU, whereas unit doors connect spaces in different FUs. Image recognition methods may be used to automatically classify doors, e.g. based on geometric or material properties. For this work, however, we assume that such missing semantic data is entered manually. The latter is expected to decrease as object recognition methods improve. In addition to being useful for space classification, capturing object data could support applications such as asset tracking.

Connectivity relations between doors, windows, furnishing elements, circulation elements, and spaces are either used directly from topology reconstruction output or they are derived from existing relations. For example, the Door-Space relation is derived from Door-Wall and Wall-Space relations, which are generated by topology reconstruction. Walls are not required directly for space classification.

5.4 Derivation of Functional Units

Window-Space and Door-Space connectivity relations in the space classification view form a graph that is termed WDS connectivity graph. Nodes correspond to doors, windows, and spaces, and edges to connectivities between them. The WDS graph is used to derive FUs and space FPs (Sect. 5.5).

The space classification procedure is illustrated with a test model of a typical floor of a multi-unit residential building (Figs. 11 and 12). The model consists of a pair of 2 bedroom residential units and a primary circulation unit with a stair case and an elevator (Fig. 12c). Spaces enclose three voids, which are duct spaces. It is assumed that no point cloud data is available for such inaccessible spaces.

The WDS connectivity graph for the test model is shown in Fig. 11a. All nodes are connected. Thus the graph has a single component. Two unit doors that provide access to the residential units from the stair case are highlighted. Unit doors are commonly the only means to access residential FUs from circulation FUs. In the WDS connectivity graph, each node that represents such a unit door is therefore a cut node. In general, the removal of cut nodes breaks a graph into multiple components. In case of the WDS connectivity graph, removal of unit door nodes breaks the graph into multiple components, where each component corresponds to an FU. In the test model, removal of the two unit door nodes breaks the WDS connectivity graph into three components that correspond to three FUs. At this point, FPs of FUs are unknown. These are determined in the next step when FPs of spaces are derived.

Table 1. Space classification: required data and their sources. Primary and secondary space functions are derived by the space classification procedure.

Object	Object property	Source		
		Topology reconstruction	Manual data input	Space classification
Space	Primary function			•
	Secondary functions			•
	Floor area	•		
Door	Object type	•		
	Object sub-type: ‘normal door’, ‘unit door’, ‘elevator door’		•	
Window	Object type	•		
Furnishing elements	Object type		•	
	Object sub-type: ‘bathroom fixture’, ‘kitchen element’, ‘laundry element’		•	
Circulation elements	Object type		•	
	Object sub-type: ‘stairs’, ‘ramp’		•	
Relations	Door-Space	derived		
	Window-Space	derived		
	Furnishing element-Space		•	
	Circulation element-Space		•	

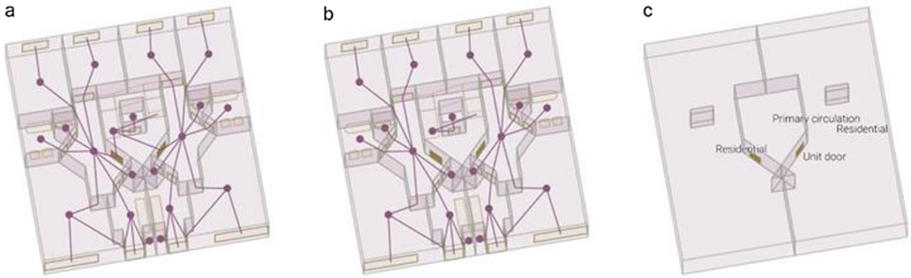


Fig. 11. Derivation of functional units (FUs) for a test model. (a) WDS connectivity graph with door, window, and space nodes. (b) WDS connectivity graph after the removal of unit door nodes. (c) Visualization of FUs obtained by merging spaces in FUs. Functional properties (FPs) of units are derived in Sect. 5.5. For clarity, gaps between spaces due to the presence of walls are not shown.

5.5 Derivation of Functional Space Properties

Heuristics are used to derive FPs of spaces. Initially, FPs of spaces are unknown (Fig. 12a). A heuristic consists of a condition that is applied to space, window, or door properties. If a space matches the condition, then one or more FPs are added to it.

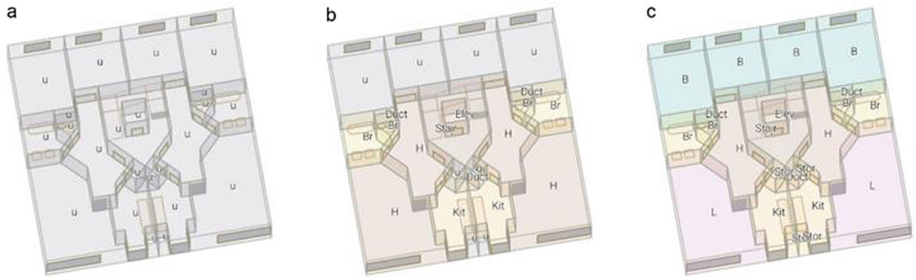


Fig. 12. Space classification steps for the test model. (a) FPs are unknown before step 1. (b) Selected FPs after step 2. (c) Primary FPs after step 4.

FPs are derived in four steps. In each step, FPs are added to spaces by evaluating heuristics. While some heuristics are specific to residential spaces, others are applicable to spaces found in other types of buildings. An example for the former is a heuristic to classify living rooms, and for the latter a heuristic to classify hallways. For the sake of simplicity, heuristics cover common cases only. That is, FPs derived by these heuristics may need to be overridden manually in exceptional cases.

In the first step, spaces are evaluated individually. Quantities and sizes of windows, doors, or other elements in a space are considered. For example, a space is considered as a hallway if it has more than two doors, or as a bathroom if it has at least one bathroom element. Similarly, a space with stairs is considered as a stair case. Elevator doors are used to determine elevator spaces. An elevator door typically connects an elevator space with another circulation space, such as a hallway or a stair case. If two spaces are related by an elevator door, then the space with the smaller floor area is considered as an elevator space. As elevator spaces span multiple floors, the space with the greater ceiling height may be considered as elevator space. However, this is insufficient if an elevator space is modeled as multiple elevator spaces with ceiling heights of floors.

FPs of FUs are determined in the second step. This helps resolve unknown FPs in the third step. FUs that have at least one kitchen space (as determined in the first step) are residential units, and primary circulation units otherwise. In primary circulation units, FPs for primary circulation are added to stair cases, elevators, and hallways. Thereby, they are designated as part of the common circulation system, as opposed to circulation spaces in residential FUs, which are private spaces. Space classification for primary circulation spaces is complete after this step. In the test model, two FUs are classified as residential units, and one FU as a primary circulation unit (Fig. 11c).

In the third step, FPs of spaces in residential units are determined that are still unknown or incomplete. In the test model, there are four spaces with unknown FPs (Fig. 12b). Moreover, the two living rooms are currently classified as hallways. The approach is to derive FPs by comparing floor areas of spaces in each unit. Heuristics in this step cover FPs for the most common space types in multi-unit residential buildings. These include living rooms, bedrooms, and storage spaces.

The largest space in a unit is considered as a living room. Spaces with access to daylight and a large enough floor area are classified as bedrooms. As the floor area for bedrooms may vary considerably among buildings or residential units in a building, using an absolute range to distinguish bedrooms from smaller spaces is error prone. To address this issue, floor areas are mapped to a qualitative floor area metric. Areas are first ordered. Then categories ‘small’, ‘medium’, and ‘large’ are determined based on the largest differences between subsequent areas. Unless they are already classified as living room, spaces with ‘medium’ or ‘large’ area and access to daylight are classified as bedrooms, whereas ‘small’ spaces and unlit spaces are classified as storage spaces.

In the fourth step, primary FPs are derived from FPs computed in the previous steps. This is done in two iterations for each space. Heuristics for circulation spaces are evaluated before those for residential spaces. For example, a space with a ‘hallway’ FP may initially have ‘hallway’ as its primary FP. Subsequently, if it has also a ‘living room’ FP, the ‘hallway’ primary FP is replaced by ‘living room’. Figure 12c shows primary FPs for spaces in the test model. Two living rooms and four bedrooms have been derived in steps 3 and 4.

6 Conclusions and Outlook

Reconstruction of information-rich as-built building models from point clouds is a challenging problem. In this paper, we have described a data processing pipeline. Starting with registered point cloud data as input, the pipeline detects first surfaces, then produces a topologically correct as-built model, which subsequently is enriched with space classification data. In this section, we discuss the limitations of our approach and provide an outlook for future work.

Point cloud data processing is currently limited to planar surfaces. We plan to extend it to the reconstruction of non-planar surfaces, for example, using NURB based RANSAC. This is relevant in case of older or historic buildings which frequently have curved features. Moreover, this would help modeling surfaces which are slightly non-planar in the real world.

A promising direction for future work concerns the recognition and modeling of built-in building components, such as windows as well as furnishing and circulation elements. In addition to space and building component data, these objects are required for space classification. Currently, data about these objects are modeled manually. With image recognition methods maturing rapidly, it is conceivable to detect them as part of point cloud data processing instead of assuming manual input of their data. The described topology reconstruction is easily extended to derive rich topological relations for built-in components and other elements.

At present time, topology reconstruction is restricted to objects with planar surfaces. A boundary representation for the initial geometry of each building component is a set of planar triangles. The extension to curved surfaces is a challenge and part of future work.

Another field of further investigations in topology reconstruction is also a deep analysis of the required run time behavior and the storage requirement. In principle, parts of reconstruction can be parallelized. Investigations are necessary to benefit from computers with several processors in this field.

Classification of dining rooms, or spaces common in upscale condominiums, such as family rooms or libraries, is currently not supported by space classification. More elaborate analysis e.g. of the access structure, or comparison with similar buildings with known FPs may help expand this step to more diverse space types.

The space classification procedure relies on heuristics to determine functional space properties. Heuristics are invoked explicitly in each step of the procedure. The current prototypical implementation of space classification in a procedural programming environment is limited. Rule-based programming methods may be more flexible to model space classification logic and control its application.

In summary, we have described a general concept for the proposed processing pipeline for reconstruction of as-built building models. We have outlined data processing and algorithms as well as required data for each step of the pipeline. Further development will involve detailed elaboration of the pipeline, validating and expanding its applicability in real-world buildings.

References

ANSI/BOMA: Gross areas of a building: standard methods of measurement (2010)

ANSI/BOMA: Multi-unit residential buildings: standard methods of measurement (2010)

Baumgart, B.G.: Winged Edge Polyhedron Representation. Stanford University, Stanford, CA, USA (1972)

Borrmann, A., Rank, E.: Specification and implementation of directional operators in a 3D spatial query language for building information models. Adv. Eng. Inf. **23**, 32–44 (2009a)

Borrmann, A., Rank, E.: Topological analysis of 3D building models using a spatial query language. Adv. Eng. Inf. **23**, 370–385 (2009b)

Borrmann, A., Schraufstetter, S., Rank, E.: Metric operators of a spetial query lanuage for 3D building models: octree and B-Rep approaches. J. Comput. Civ. Eng. **23**, 33–46 (2009c)

CSI: The Construction Specifications Institute: OmniClass Construction Classification System (OCCS), Table 13: Spaces by Function (2010)

DIN: DIN 277-1:2016-01, Grundflächen und Rauminhalte im Bauwesen (2016)

DIN: DIN 283-2:2004-2, Wohnungen - Teil 2: Berechnung der Wohnflächen und Nutzflächen (2004)

Egenhofer, M.; Herring, J.: A mathematical framework for the definition of topological relationships. In: Proceedings of the 4th International Symposium on Spatial Data Handling, pp. 803–813 (1990)

GSA: BIM Guide for 3d Imaging (2009)

Huhnt, W.: Reconstruction of edges in digital building models. Advanced Engineering Informatics, (2018, in review)

Jung, J., Hong, S., Jeong, S., Kim, S., Cho, H., Hong, S., Heo, J.: Productive modeling for development of as-built BIM of existing indoor structures. Autom. Constr.**42**, pp. 68–77 (2014)

Kraft, B.: Ein Verfahren der Raumzerlegung als Grundlage zur Prüfung von Geometrie und Topologie digitaler Bauwerksmodelle, Doctoral thesis, TU Berlin, Germany (2016)

Mäntylä, M.: An Introduction to Solid Modeling. Computer Science Press, Rockville (1988)

Mura, C., Mattausch, O., Villanueva, A.J., Gobbetti, E., Pajarola, R.: Automatic room detection and reconstruction in cluttered inddor environments with complex room layouts. Comput. Graph. **44**, 20–32 (2014)

Nguyen, T.-H., Oloufa, A.A., Nassar, K.: Algorithms for automated deduction of topological information. Autom. Constr. **14**, 59–70 (2005)

OSCRE: Open Standards Consortium for Real Estate: Space Classification Code List

Sietzema, A.: An exploration of COBie based point cloud processing for facility management in Qatar. Master Thesis University of Twente (2015)

Wang, C., Cho, Y., Kim, C.: Automatic BIM component extraction from point clouds of existing buildings for sustainability applications. Autom. Constr. **56**, 1–13 (2015)

Wang, M., Tseng, Y.-H.: Incremental segmentation of LIDAR point clouds with an octree-structured voxel space. Photogram. Rec. **26**, 32–57 (2011)

Xiong, X., Adan, A., Akinci, B., Huber, D.: Automatic creation of semantically rich 3D building models from laser scanner data. Autom. Constr. **31**, 325–337 (2013)

Yousefzadeh M.: As-Built modeling for energy simulation. Professional Doctorate thesis. University of Twente (2016)

State-of-Practice on As-Is Modelling of Industrial Facilities

Eva Agapaki[1(✉)] and Ioannis Brilakis[2]

[1] Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK
{ea437,ib340}@cam.ac.uk
[2] Laing O'Rourke Reader, Department of Engineering,
University of Cambridge, Cambridge, CB2 1PZ, UK

Abstract. 90% of the time needed for the conversion from point clouds to 3D models of industrial facilities is spent on geometric modelling due to the sheer number of Industrial Objects (IOs) of each plant. Hence, cost reduction is only possible by automating modelling. Our previous work has successfully identified the most frequent industrial objects which are in descending order: electrical conduit, straight pipes, circular hollow sections, elbows, channels, solid bars, I-beams, angles, flanges and valves. We modelled those on a state-of-the-art software, EdgeWise and then evaluated the performance of this software for pipeline and structural modelling. The modelling of pipelines is summarized in three basic steps: (a) automated extraction of cylinders, (b) their semantic classification and (c) manual extraction and editing of pipes. The results showed that cylinders are modelled with 75% recall and 62% precision on average. We discovered that pipes, electrical conduit and circular hollow sections require 80% of the Total Modelling Hours (TMH) of the 10 most frequent IOs to build the plant model. TMH was then compared to modelling hours in Revit and showed that 67% of pipe modelling time is saved by EdgeWise. This paper is the first to evaluate state-of-the-art industrial modelling software. These findings help in better understanding the problem and serve as the foundation for researchers who are interested in solving it.

Keywords: Industrial facilities · Facility management
Building information modelling

1 Introduction

"As-Is" Building Information Models (AI-BIMs) are the 3D digital representation of the existing condition of facilities and encompass geometric definitions at different levels of aggregation and parametric rules [1]. The clear majority of large refineries were built before the advent of CAD in 1977: as-is models, therefore, do not exist to assist their maintenance operations [2, 3]. AI-BIMs of industrial plants have substantial impact in various applications. Some of these include maintenance, strategic planning of their operations, revamping purposes, retrofitting of old sites and preparation for dismantling [4–7].

Inexistence of AI-BIMs will result in time lags for these operations. This is crucial for industrial managers, since without detailed planning, productivity will be

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 103–124, 2018.
https://doi.org/10.1007/978-3-319-91635-4_6

substantially affected, and the agreed budget and timeline expectations will not be met. Moreover, there are thresholds on the acceptable shut down duration that will not impede production, and those limits cannot be violated without incurring extra costs. For instance, [8] reported that 40% of the total 3D modelling cost of retrofitting a Chevron plant was spent on data-processing labor and the shut-down time was limited to 72 h to avoid additional costs. Every modelling hour saved can prevent critical failures or unexpected accidents, thus continuous production flow of these assets is achieved. This work aims to assist the tedious current practice in this regard.

Modelers use the following four main steps to manually process AI-BIMs: (a) data collection, (b) point cloud registration, (c) geometric modelling and (d) addition of accompanying information. Initially, data is collected using laser scanners and photogrammetry, which are represented by their cartesian or polar coordinates, the point cloud, and in some cases by their color data (RGB). The scans need to be registered in a consistent coordinate system by calculating inter-scan rigid body transformations and the registered point cloud represents the complete measured data. Then this data needs to be geometrically modelled.

Geometric modelling entails (a) primitive shape detection, (b) semantic classification of detected shapes and (c) fitting. Firstly, primitive shapes are detected (e.g., cylinders, tori, planes) and labelled (e.g., pipes, elbows, I-beams). Afterwards, the primitives are fitted to known solid shapes to obtain their geometric parameters. Their relationships to other objects need to be obtained in order to produce a complete AI-BIM in the Industry Foundation Schema (IFC) format. IFC is a data format that allows geometric, material and other construction related information to coexist in a single model.

Geometric modelling is the “bottleneck” during the Scan-to-BIM process of any industrial facility given how costly and time consuming it is. Recent studies have reported that geometric processing takes 90% of the modelling time [9, 10]. [10] reported that 10 operators were needed to process 1084 scans of a nuclear reactor and model its objects in around 6 months using Dassault Systems SolidWorks and Trimble Realworks. In contrast, laser scanning of the plant was completed in only 35 days. This significant time required to model the vast number of industrial objects impedes adoption of as-is 3D modelling for these plants.

The research presented in this paper is exploratory in nature, not causal. It does not seek to solve the problem of automating the modelling of industrial facilities. It rather seeks to improve our understanding of the problem and the extent to which it has been resolved so far and provide a foundation for future researchers interested in solving it. This is why the main objective of this paper is to identify how laborious industrial objects are for modelling, as well as to measure the performance of existing tools in modelling these particular object types. The authors used the most frequent objects based on a statistical analysis of 3D modelled industrial objects in a variety of industrial plants as explained in [11]. An overview of the state-of-the-art tools available for as-is modelling is given to select the most advanced tool for evaluation. The most frequent objects were modelled in the state-of-the-art, semi-automated modelling software, EdgeWise, and their average modelling time was measured. The level of automation of EdgeWise is also measured for the most frequent industrial object types. This analysis will substantially assist automated modelling efforts to efficiently reduce modelling time and facilitate facility management.

2 Background

Industrial plants can be divided into ten main categories [12]: (a) onshore and (b) offshore oil platforms, (c) chemical, (d) mining, (e) pharmaceutical plants, (f) power plants, (g) water and wastewater treatment facilities, (h) natural gas processing and biochemical plants, (i) refineries, (j) food processing factories, (k) defense facilities, (l) metal production facilities, (m) nuclear plants, (n) research facilities and (o) warehouses and silos. The object types of industrial facilities belong to the main object categories: (a) structural elements, (b) piping system, (c) electrical, (d) safety and (e) general equipment, (f) architectural elements, (g) instrumentation, (h) Heating, Ventilation and Air Conditioning (HVAC) and (i) civil elements.

The most frequent object categories being around 90% of all objects in these facilities are: structural elements (33%), the piping system (28%) and electrical equipment (27%) based on our previous work [11]. The most frequent object types of these categories are in descending order: electrical conduit (24%), straight pipes (15%), circular hollow sections (6.5%), elbows (5.4%), channels (5%), solid bars (4.5%), I-beams (4.4%), angles (4%), flanges (3.3%) and valves (2%) as presented in the same work.

2.1 Automated Industrial Plant Modelling

State-of-the-Art Software

Almost all available modelling tools of industrial objects depend on human intervention for most of the modelling tasks. Leading 3D CAD software (Autodesk, Bentley, AVEVA and FARO) have developed programs containing a variety of functions that enable 3D plant modelling and visualization from 3D point clouds. For example, AutoCAD Plant 3D accompanied with FARO's PointSense Plant add-in enables semi-automated pipe modelling from Point Clouds. PointSense Plant provides several functions and a large standard library with a variety of piping and structural components available for the detection of pipelines from 3D point clouds. Moreover, fitting template objects to scanned 3D objects is performed automatically and constraints can be applied to fix potential errors of fitting. PointSense Plant 17.5 has integrated a pre-calculation tool that detects cylinders in the point cloud of a specific area and has the ability to colorize the Point Cloud by deviation from reference geometry [13]. However, the users still manually model the as-is pipelines by finding the insertion points for fitting CAD objects to the segmented 3D point clouds and fitting errors of the extracted cylinders are not provided. The "Walk the Run" feature is rather a suggestion for pipe insertion points than an automated pipe modelling tool.

EdgeWise is another semi-automated platform that is extensively used. The main difference of Pointsense and EdgeWise is that a modeler using the former should extract the desired boundaries of an object manually and afterwards the software will automatically extract the correct dimensions and location. However, this procedure is automatically performed by EdgeWise, that is why it was chosen as the most suitable tool for evaluation of the most frequent industrial object types that will be presented in the next section. Structural sections are modelled manually in all available software packages. Fitting of user-selected primitives (e.g., cylinders, cuboids, tori etc.) is performed automatically by both EdgeWise and PointSense Plant. To date, no one has provided viable assessments of state-of-the-art tools.

The details behind the algorithmic development of software are commercially protected (trade secrets), so we can evaluate only the outcomes and not the algorithms that are used to reach those outcomes.

State-of-Research

State-of-the-art research work on pipe detection has partially solved the problem and not to a greater extent compared to commercially available software like EdgeWise [14, 15]. For instance, [14] only detect pipes in orthogonal directions. A recent study completed by [15] is dependent on threshold values for radius and normal estimation. The pipe radius range is 0.0254 m–0.762 m and the normal deviation is 5°. Therefore, [15] cannot be generalized for pipe detection. Their updated Hough Transform based on [4] study detects pipes in two sample datasets with 60% recall and 89% precision.

Prior knowledge of industrial scenes has assisted researchers to detect industrial objects. [16] used prior knowledge (Piping and Instrumentation Diagram, P&ID) to detect Mechanical, Electrical and Plumbing equipment (MEP). However, as-is P&IDs are often not available as prior knowledge in industrial plants, thus they do not reflect the modifications a plant undergoes through its life. For this reason, prior knowledge cannot rely on P&IDs. [17] used topological information to extract semantic labels for four object classes: pipes, planes, elbows and valves. They detect cylinders with 86% precision and 92% recall. However, their semantic labels consider that all cylindrical objects are pipes, without investigating other potential object classes with the same shape.

2.2 Gaps in Knowledge and Research Questions

Considering the state of practice and body of research reviewed above, existing studies for as-is modelling of industrial plants have focused on automated detection of cylindrical objects and no scientific and viable evaluation of existing state-of-the-art software tools is provided. It is therefore still unclear (1) how much time it takes to model those in state of the art software and (2) the level of automation achieved with state of the art software.

The aim of this work is to solve the gaps in knowledge by answering the following research questions:

(a) What is the time required for modelling the most frequent object types in state-of-the-art software?
(b) How can state-of-the-art as-is modelling tools be assessed in terms of automated detection of objects achieved?

3 Research Methodology Framework

The research conducted in this paper is exploratory in nature. The most frequent object types as derived in [11] are modelled in EdgeWise to measure the modelling time of each type and the performance of the software is evaluated. The time required for manual modelling of cylindrical objects was then compared with that measured in EdgeWise. Research efforts on automated cylinder extraction are then investigated to compare the evaluation results from EdgeWise and set the ground for future research towards minimizing the modelling time of these assets.

3.1 Data Collection and Assumptions

Four case studies of laser scanned industrial facilities were examined to have a representative sample of industrial objects in different facilities. Two case studies are rooms of an industrial plant, one was a water treatment facility in Cambridge and the fourth was a room of a petrochemical plant. The industrial and petrochemical plant are anonymized since rights are reserved by AVEVA Group Plc. The scanner setup and scan frequency of these facilities is not available, since data was collected by industrial partners. The water treatment facility in Cambridge was laser scanned by the authors. We used a Faro Focus 3D X330 laser scanner to collect 6 laser scans with resolution of at least 1 point/cm^2 and ranging error ±2 mm.

The average frequencies of industrial objects used in this paper were taken from the results of [11] and represent the average frequency of appearance of 3D modelled datasets investigated. Solid bars are not modelled separately in EdgeWise since they cannot be distinguished from circular hollow sections in a laser survey.

3.2 EdgeWise Evaluation for Pipeline Modelling

Four sample point cloud datasets were used to evaluate the capabilities of EdgeWise and obtain modelling times for the most frequent object types. Figure 1 shows the sample datasets that were used for this evaluation.

Pipeline modelling is significantly assisted by the automated extraction of cylinders that EdgeWise provides. The scans were processed on a desktop computer with CPU Intel® Core^{TM} i7-4790K at 4.00 GHz, 32 GB RAM and Windows 10 64-bit operating system. The average processing time for this operation using the above-mentioned operating system for the sample datasets is $3.3 * 10^{-3}$ min/ (cylinder*points in the point cloud), as shown in Table 3. The average number of points of all datasets used is 258 million and the number of points of each dataset is presented in Table 1. The average diameter of cylinders and pipes is presented for evaluation purposes in Table 2.

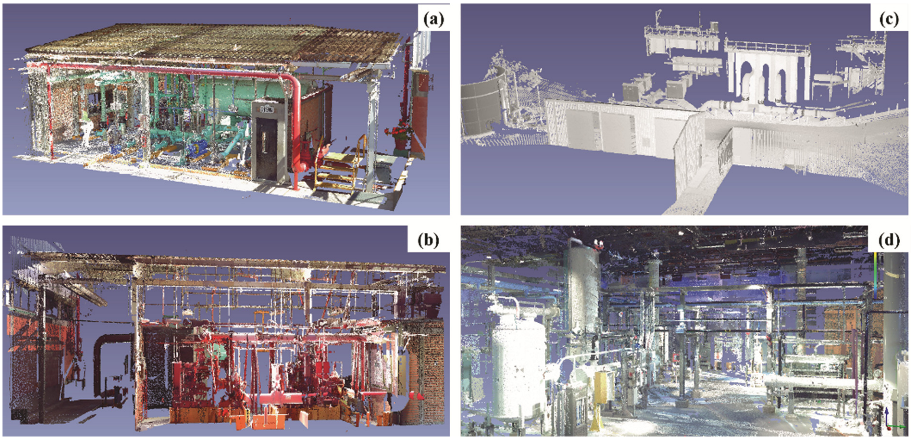


Fig. 1. Sample datasets for evaluation. (a), (b) Two rooms of a typical industrial facility, (c) a water treatment facility and (d) a room of a petrochemical plant

Table 1. Total number of points in the point cloud datasets and total number of cylinders and pipes in each case study

	Typical facility Room 1	Typical facility Room 2	Water facility	Petrochemical plant
	Total number			
Points (millions)	129	105	122	675
Automatically detected cylinders	551	86	44	358
Manually detected cylinders	166	79	48	265

Table 2. Average diameter of cylinders and pipes for each dataset

	Typical facility Room 1	Typical facility Room 2	Water facility	Petrochemical plant
	Average diameter (m)			
Cylinder	0.067	0.076	0.315	0.095
Pipe	0.114	0.106	0.617	0.081

We set the parameters used for cylinder extraction to a minimum of 80 points, in order to detect a pipe and provide a distance tolerance to 0.7 * 10^{-3} m. The minimum threshold of the software is 50 points to identify pipelines, however if we give a very low value, the automated extraction tool will identify noisy and erroneous features as pipes. The distance tolerance is a parameter that determines how far away from the cylinder a 3D point can be, so that it is not excluded from the extraction algorithms. The default value of 0.7 * 10^{-3} m is used here, which was obtained from a scanner with a high level of accuracy and low noise [18].

After the automated extraction step, the cylinders were inspected and approved depending on the modeler's discretion (classification). For cases where it was difficult to identify the object, pictures taken from the laser scanner were used to assist the inspection process. A user friendly "*Smart Sheet*" was produced, which contains information such as the length, diameter, Root Mean Square Error (RMSE) and coverage (%) of each pipe spool. The results show that although cylinders are automatically extracted, no contextual information is provided. Henceforth, electrical conduit, handrails, cylindrical pipe supports, vessels and other object types were modelled as straight pipes.

The next step in the evaluation process was to edit the pipes and to manually add missing ones. Using the "*Easy Connect*" tool pipe spools were connected, and tees and elbows were added in the piping network. Then, labels were manually assigned for each cylinder that was automatically extracted by the software and metrics were used to evaluate the software's performance.

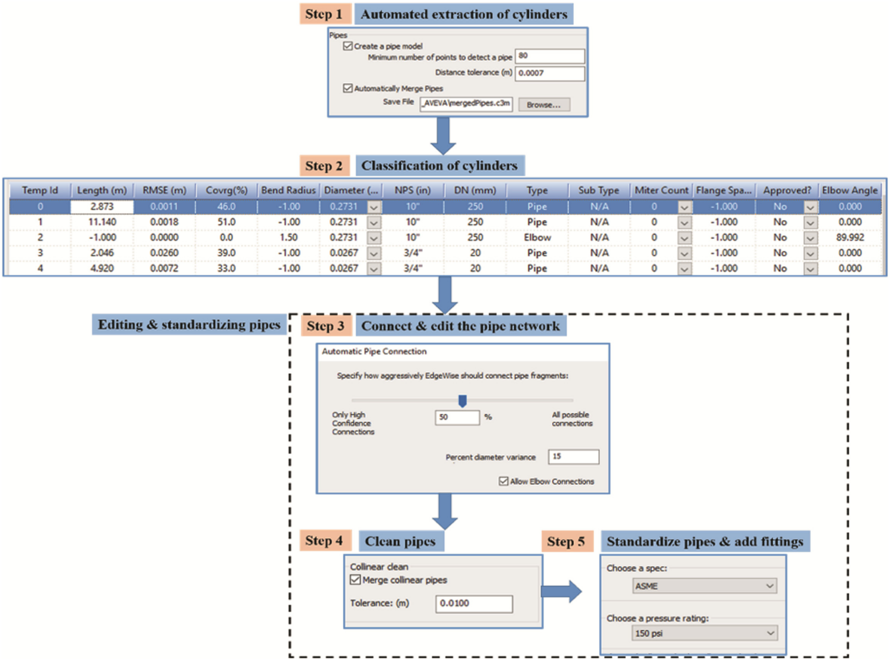
An additional step of cleaning the pipes and merging the connecting spools together was performed to complete the pipeline system. This step was completed automatically by the software. Then, standard catalogues were used to get standardized pipe dimensions. We chose the American Society of Mechanical Engineers' (ASME) specifications and pressure rating of 150 psi for our datasets. After this step, fittings, such as flanges and valves, were applied on the standardized pipes. There are different types of standard fittings that the user can select from available standard libraries.

Table 3. Modelling time needed for each modelling task for each dataset and average time per object (min/object)

Modelling task	Typical facility Room 1	Typical facility Room 2	Water facility	Petrochemical plant	Average time (min)
	Time (min)				
Automated extraction of cylinders[a]	$1.5 * 10^{-3}$	$1.5 * 10^{-3}$	$7.1 * 10^{-3}$	$1.4 * 10^{-3}$	$3.3 * 10^{-3}$
Semantic classification of cylinders[b]	0.20	0.47	0.17	0.12	0.24
Manual extraction & editing of pipes[c]	0.69	2.37	2.43	1.22	1.68

[a]per cylinder * point, [b]per cylinder, [c]per pipe

The modelling of pipelines is summarized in three basic steps: (a) automated extraction of cylinders, (b) semantic classification of cylinders and (c) manual extraction and editing of pipes. Fitting is performed automatically during manual or automated extraction: therefore, it is not a separate step of the procedure. The steps of the procedure for pipeline modelling in EdgeWise are presented in Fig. 2.


Step 1
Automated extraction of cylinders
Step 2
Classification of cylinders
Editing & standardizing pipes
Step 3
Connect & edit the pipe network
Step 4
Clean pipes
Step 5
Standardize pipes & add fittings

Pipes
☑ Create a pipe model
Minimum number of points to detect a pipe 80
Distance tolerance (m) 0.0007
☑ Automatically Merge Pipes
Save File _AVEVA\mergedPipes.c3m Browse...

Temp Id	Length (m)	RMSE (m)	Covrg(%)	Bend Radius	Diameter (...	NPS (in)	DN (mm)	Type	Sub Type	Miter Count	Flange Spa...	Approved?	Elbow Angle
0	2.873	0.0011	46.0	-1.00	0.2731	10"	250	Pipe	N/A	0	-1.000	No	0.000
1	11.140	0.0018	51.0	-1.00	0.2731	10"	250	Pipe	N/A	0	-1.000	No	0.000
2	-1.000	0.0000	0.0	1.50	0.2731	10"	250	Elbow	N/A	0	-1.000	No	89.992
3	2.046	0.0260	39.0	-1.00	0.0267	3/4"	20	Pipe	N/A	0	-1.000	No	0.000
4	4.920	0.0072	33.0	-1.00	0.0267	3/4"	20	Pipe	N/A	0	-1.000	No	0.000

Automatic Pipe Connection
Specify how aggressively EdgeWise should connect pipe fragments:
Only High Confidence Connections 50 % All possible connections
Percent diameter variance 15
☑ Allow Elbow Connections

Collinear clean
☑ Merge collinear pipes
Tolerance: (m) 0.0100

Choose a spec: ASME
Choose a pressure rating: 150 psi

Fig. 2. Workflow of pipeline detection, classification and editing steps in EdgeWise

The average processing time per cylinder or pipe for each step is computed in Table 3 and calculated as following:

$$\text{Time/cylinder.point} = \text{TAE/AC} * \text{P} \tag{1}$$

Where TAE is the time for automated extraction of cylinders, AC is the number of automatically detected cylinders and P is the number of points in the dataset

$$\text{Time/pipe} = \text{TME/TP} \tag{2}$$

Where TME is the time for manual extraction and editing of pipes and TP is the total number of pipes

$$\text{Time/cylinder} = \text{TC/AC} \tag{3}$$

Where TC is the time for semantic classification of cylinders and AC is the number of automatically detected cylinders.

The number of automatically detected cylinders and the total number of pipes is shown in Table 1 for each case study. The latter is the sum of automatically and manually detected pipes in each dataset. These normalizations are used to compare the modelling times for each case study, since the number of points and cylinders processed are different for each dataset.

The time for semantic classification was 0.24 min per cylinder on average. Manual extraction and editing of pipes was the most time-intensive step, since we needed 1.68 min per pipe on average to manually add missing pipes and edit the existing ones. The observations show that the manual effort to classify and extract pipes was 1.92 min per cylinder on average, which is the summation of two subsequent steps, semantic classification of cylinders and manual extraction and editing of pipes. This is almost three times the time needed for automated extraction of cylinders by the software.

A variation of the time needed for automated extraction of cylinders between the water facility and the other datasets is observed. This discrepancy is attributed to the fact that the water facility is an outdoor facility, requiring the most processing time compared to the other datasets. Technically, outdoor scenes are inherently more occluded and incomplete exhibiting extreme variations in point density [19]. These effects are mitigated by the limited size and constrained shape of rooms. The two rooms of the typical industrial facility were processed at the same time in our operating system, for this reason the time required for automated extraction is the same as shown in Table 3. Manual modelling of the second room of this facility required the most modelling time compared to the first room. This is due to cluttered pipelines, which resulted in the largest Room Mean Square Error (RMSE) of the cylinder diameters, as shown in Table 4. This clutter is attributed to the reflective surface of pipelines. Manual extraction and editing of pipes in the water facility is another modelling time outlier. Highly occluded pipelines are the primary reason for this outlier, since they have the lowest average coverage (26.5%), compared to the other projects. The diameter of pipelines in the water facility was significantly larger, since most pipes are used for sewage purposes. These observations show that manually detected pipes have larger average diameter (0.617 m) compared to automatically extracted cylinders (0.315 m) for the same dataset. This means that it is difficult for the software to identify cylinders with large diameters.

Table 4. Root Mean Square Error (RMSE) of the radius and coverage (%) of automatically detected cylinders in each dataset and average values

Automatically detected cylinders in:	RMSE of the cylinder radius (m)	Coverage (%)
Typical facility – Room 1	$1.7 * 10^{-3}$	32.5
Typical facility – Room 2	$6.7 * 10^{-3}$	30.2
Water facility	$1.9 * 10^{-3}$	26.5
Petrochemical plant	$4.2 * 10^{-3}$	27.6
Average	$3.6 * 10^{-3}$	29.2

RMSE and coverage percentages for each extracted cylinder are calculated in the "*SmartSheet*", provided in EdgeWise. Table 4 summarizes their average values for all case studies. The results show that the first room of the typical industrial facility has the lowest RMSE, meaning that the extracted cylinders fit well the corresponding points of the cylinders. The average coverage area of cylinders in all case studies is around a quarter of the cylinder (29.20%), which is the reason that many cylinders are not automatically extracted.

The performance of the software is evaluated based on the two-following metrics, precision and recall [20],

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{4}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{5}$$

Where TP are the number of objects that are automatically detected as pipes and were correctly inspected as pipes, FP are the number of objects that are detected as pipes, but we classified them as other cylindrical objects (for instance handrails, circular hollow steel sections to name a few) and FN are the number of objects that are pipes but were not automatically detected as pipes. Those pipes were manually extracted and added to the model.

The performance metrics obtained from our four sample datasets are given in Table 5.

Table 5. Average performance metrics of pipe and cylinder detection

Dataset	Pipe detection metrics		Cylinder detection metrics	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
Typical facility Room 1	80.1	27.9	69.3	48.2
Typical facility Room 2	59.5	54.6	100.0	22.0
Water facility	33.3	36.4	87.3	86.4
Petrochemical plant	59.6	69.3	45.7	91.9
Average	58.1	47.0	75.6	62.1

According to precision, out of all the automatically detected cylinders only an average of 47% in all case studies correspond to pipes, whereas the rest were other cylindrical objects. The average recall was 58.1%, meaning that only 58.1% of all pipes existing in a typical facility will be automatically detected. The results show that the water treatment facility, which is an outdoors facility, has the lowest recall, being 33.3%. The low performance metrics of this dataset, compared to the other ones, can be attributed to increased noise. The low precision of pipes in the first room of the typical facility (27.9%) is attributed to a larger number of FPs (roof tiles), which were wrongly detected as pipes.

The same metrics were measured for cylinders. The only difference in the metrics used is that precision is defined as the number of automatically detected cylinders out of all the detected cylinders, whereas recall is the number of automatically detected cylinders out of all other automatically detected non-cylindrical shapes. The recall of cylinders is high for all datasets except the petrochemical plant (45.7%), which is attributed to low scan completeness of this dataset and increased clutter. The average recall for the four datasets is 75.6% indicating the advantage of the software to extract this primitive shape. The precision of cylinders is also 15% higher compared to that of pipes, since the software is designed to detect cylindrical shapes. The lowest precision (22%) is observed for the second room of the typical industrial facility, which is attributed to corrugated shapes in the roof that were incorrectly modelled as cylinders. The same trend (low precision of about 48%) is observed for the first room of the facility for the same reason.

Representative 3D models obtained from the room of the petrochemical plant, two rooms of a typical industrial facility and the water treatment plant are presented in Figs. 3, 4, 5 and 6. The initial point cloud, the automated pipeline extraction output and the final 3D model that was obtained after manual modelling of the most frequent pipeline elements, structural sections and electrical conduit are presented in the same Figures. These 3D models are not the complete 3D models of the facilities, but the subsets used for the evaluation purposes of this paper.

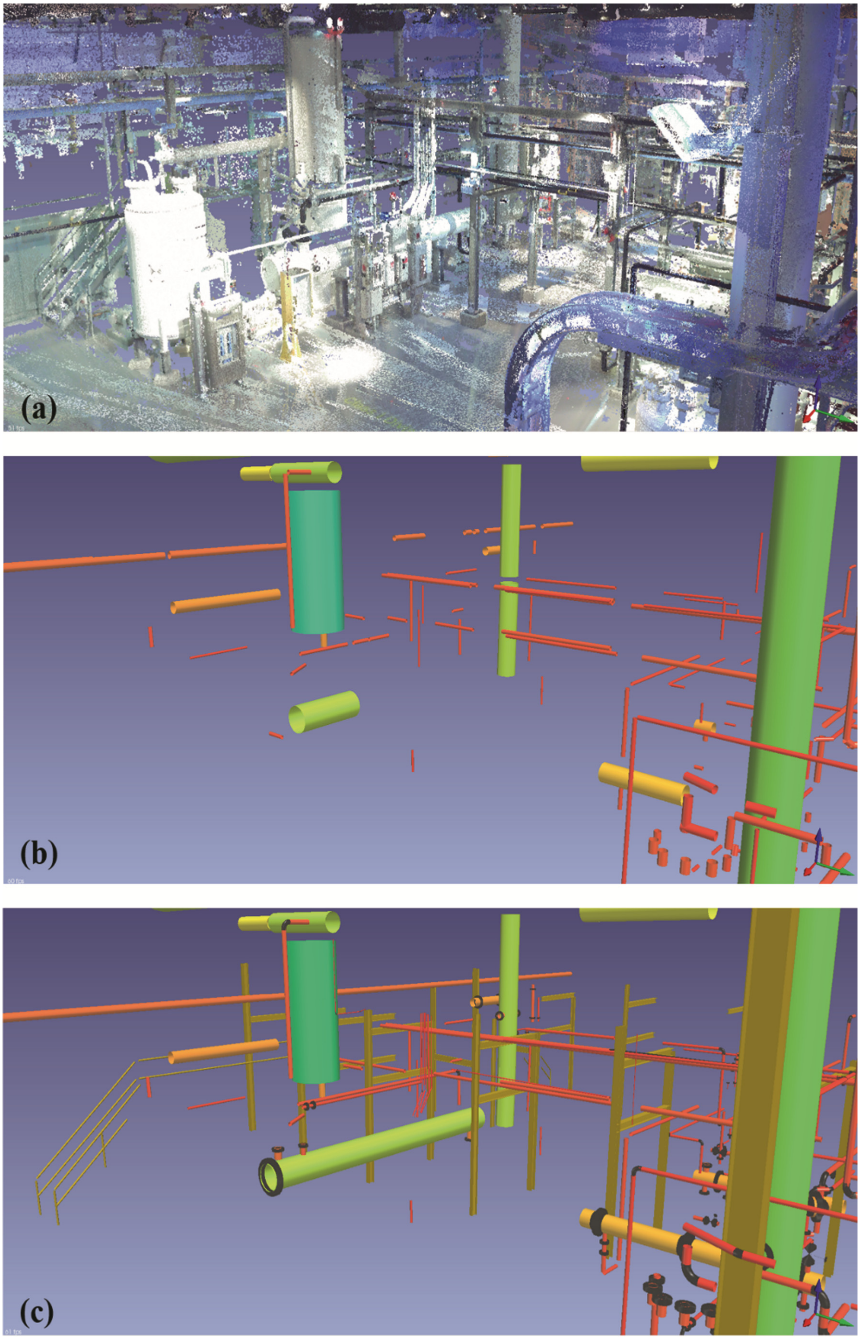


Fig. 3. Input point cloud, (b) automated cylinder extraction in EdgeWise Plant/MEP and (c) 3D model after manual modelling of pipes, structural elements and electrical conduit for a room of a petrochemical plant (dataset provided by AVEVA Group Plc.).

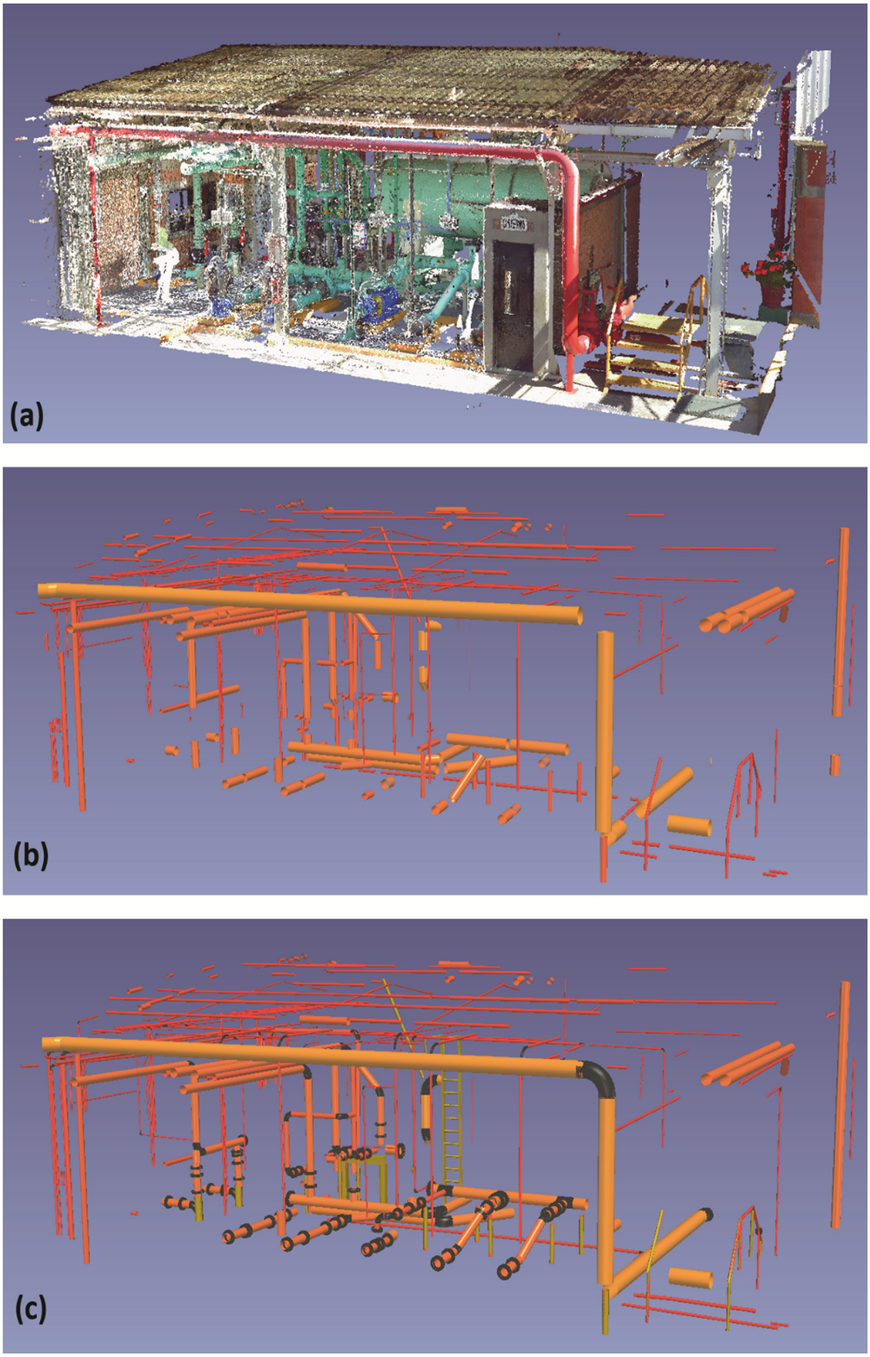


Fig. 4. Input point cloud, (b) automated cylinder extraction in EdgeWise Plant/MEP and (c) 3D model after pipe, structural and electrical conduit extraction for the first room of an industrial facility (dataset provided by AVEVA Group Plc.)

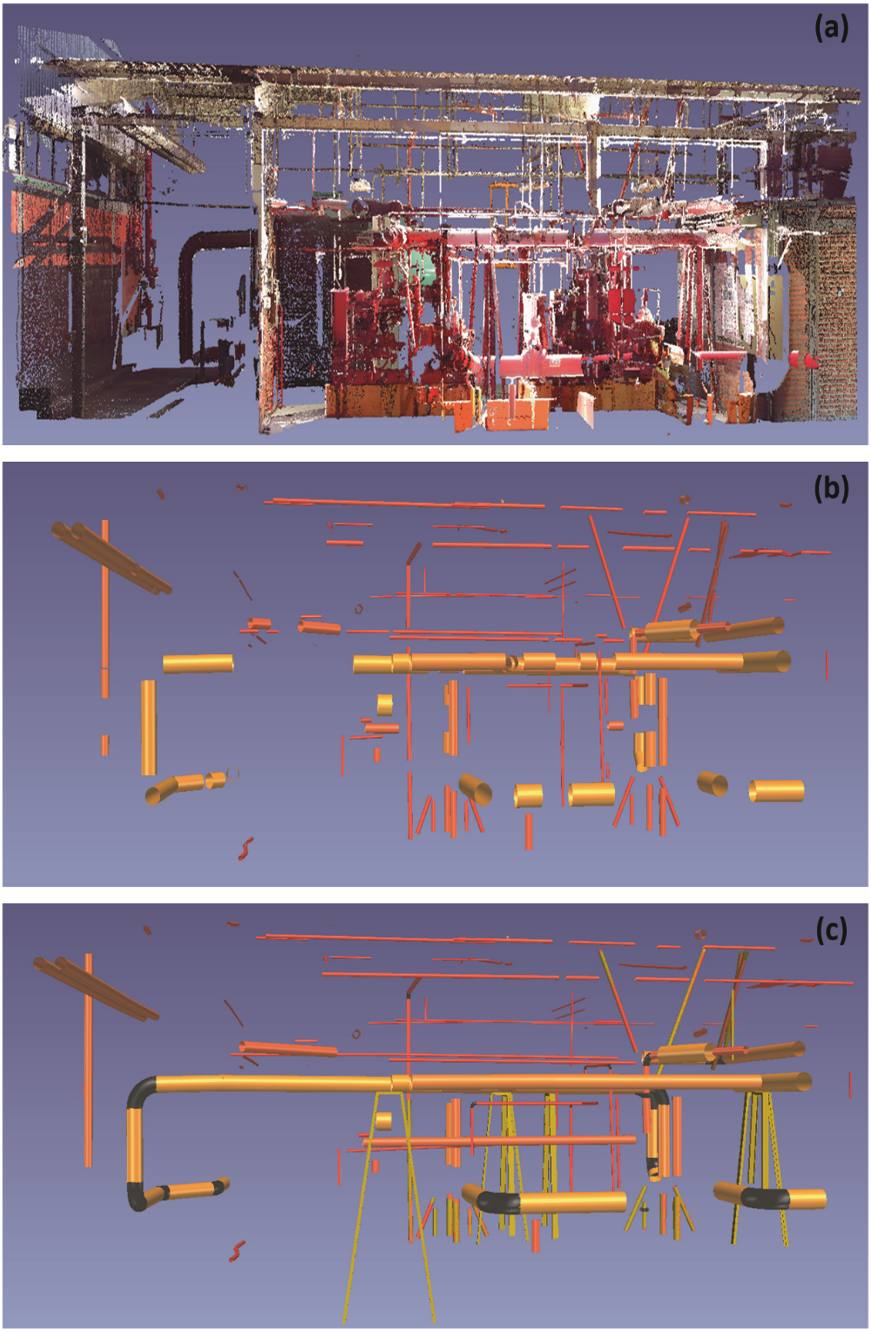


Fig. 5. Input point cloud, (b) automated cylinder extraction in EdgeWise Plant/MEP and (c) 3D model after pipe, structural and electrical conduit extraction for the second room of an industrial facility (dataset provided by AVEVA Group Plc.)

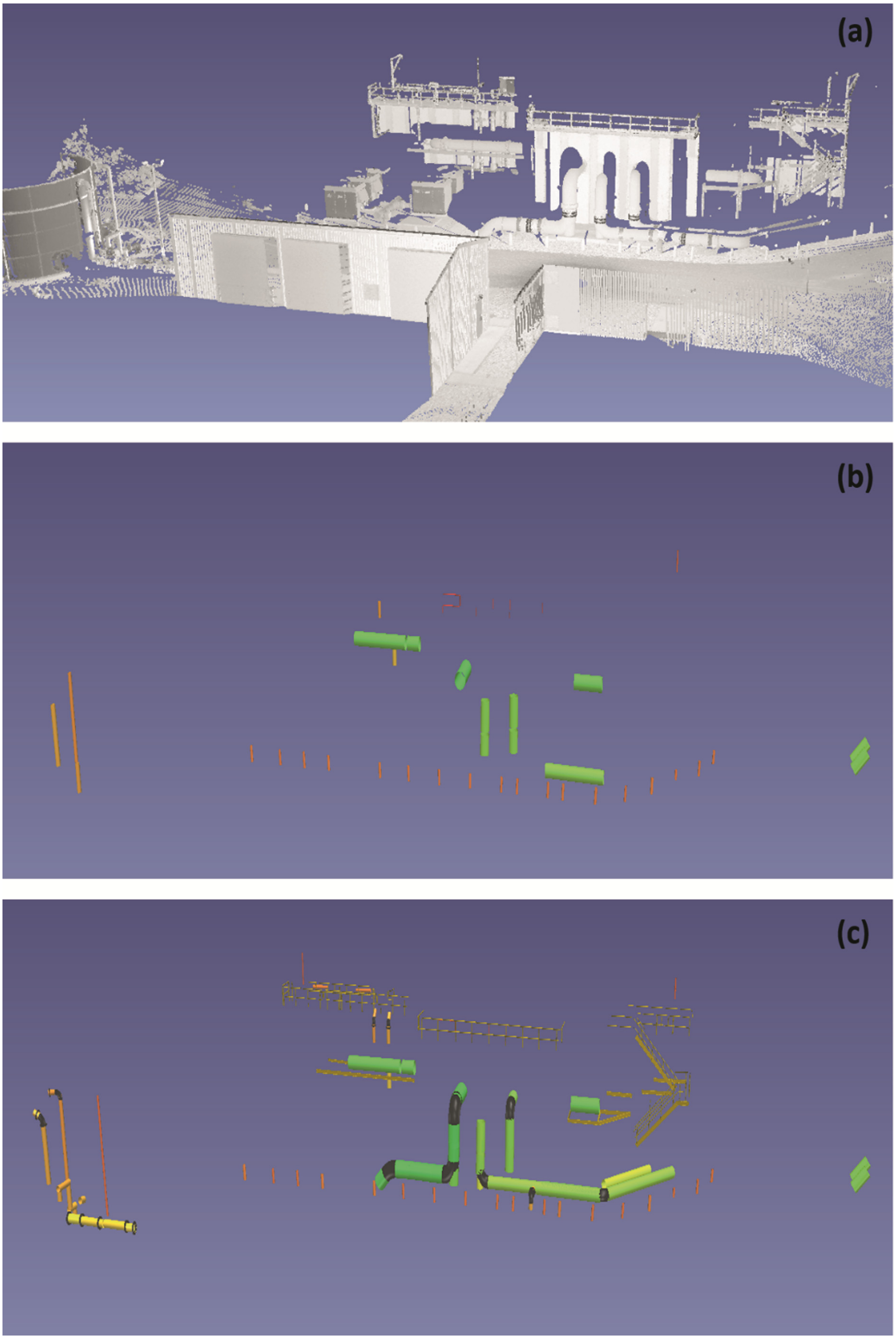
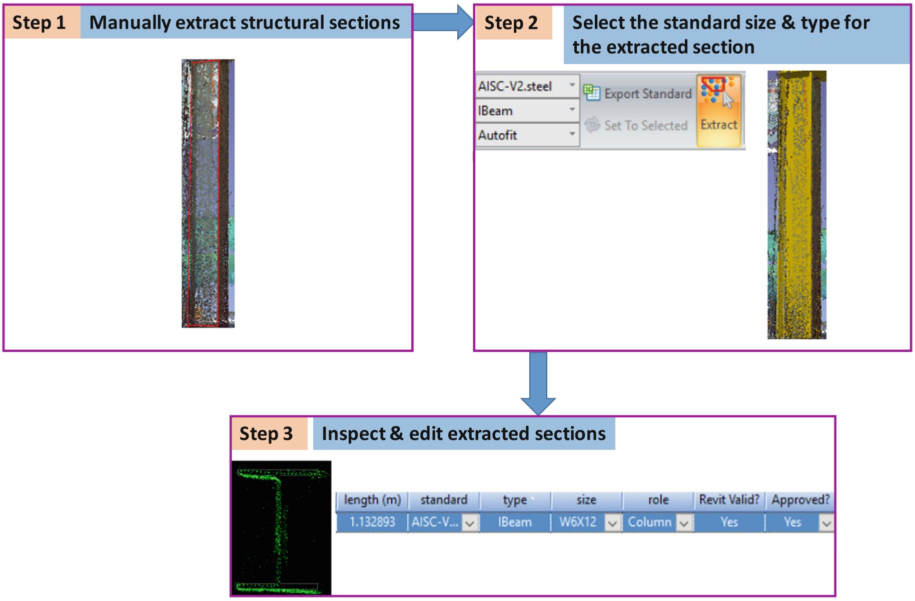


Fig. 6. Input point cloud, (b) automated cylinder extraction in EdgeWise Point/MEP and (c) 3D model after pipe, structural and electrical conduit extraction for a water treatment facility in Cambridge (dataset acquired by the authors)

3.3 EdgeWise Evaluation for Modelling of Structural Components

Software packages used for extraction of structural elements have been developed by ClearEdge3D (2017). EdgeWise Structural is used for our evaluation in this work. The most frequent structural elements that were identified in our previous work [11] (circular hollow sections, channels, solid bars, I-beams) are modelled in the four case studies.

The user selects the I-beam, Channel and RoundTubing tools to manually extract the respective elements. The user can also create custom standards for shapes that do not exist on the standards list. The "*Pattern Extract*" tools extract groups on repeatable elements of the same object type. The extracted sections are then inspected for accuracy in the "*SmartSheet*". The workflow of the manually modelled structural sections is summarized in Fig. 7.



Step 1 Manually extract structural sections

Step 2 Select the standard size & type for the extracted section

AISC-V2.steel
IBeam
Autofit

Step 3 Inspect & edit extracted sections

length (m)	standard	type	size	role	Revit Valid?	Approved?
1.132893	AISC-V...	IBeam	W6X12	Column	Yes	Yes

Fig. 7. Workflow of the manual modelling of structural elements in EdgeWise Structural.

The standards that were used for this evaluation were taken from the American Institute of Steel Construction (AISC) manuals. The authors also used the "*Autofit*" tool to extract the correct size of the specified section automatically. Precision and recall metrics were not used herein, since the procedure is manual.

3.4 Overall Performance of State-of-the-Art Modelling Software

The performance of state-of-the-art modelling software is summarized in Table 6. This Table shows that fitting of the most frequent object types has been solved by commercial software like EdgeWise, since known geometric shapes are automatically fitted to the selected point clusters. Automated primitive shape detection of cylinders has partially been solved since the results showed 75% recall and 62% precision in EdgeWise. Non-cylindrical shapes are manually extracted, and classification of all object types has not been achieved.

Table 6. Performance of state-of-the-art software packages on each modelling step for the most frequent object types

Industrial object type	Primitive shape extraction	Semantic labelling (classification)	Fitting
Straight pipe	Partially solved	Not solved	Solved
CHS[a]	Partially solved	Not solved	Solved
Channel	Not solved	Not solved	Solved
Conduit	Partially solved	Not solved	Solved
I-beam	Not solved	Not solved	Solved
Valve	Not solved	Not solved	Solved
Elbow	Not solved	Not solved	Solved
Flange	Not solved	Not solved	Solved
Angle	Not solved	Not solved	Solved

[a]Circular Hollow Section (CHS)

The 3D models can be exported to Revit, in order to obtain IFC models for interoperability purposes between different software packages. However, we observed that reducers, valves, flanges, angles and some channels (C3 and C4 according to the American Institute of Steel Construction standards - AISC) cannot be exported in Revit. Models containing straight elements with length less than 4 mm cannot also be transferred to Revit.

Pipes, conduit and Circular Hollow Sections (CHSs) were also modelled manually in Revit to compare the man-hours needed for their shape extraction through this manual process. 30 objects were modelled in each category and their average modelling times were measured. The workflow of manual modelling in a software such as Revit entails three steps: (a) manual segmentation of the desired object in a point cloud visualization software such as CloudCompare, (b) export of the points in Autodesk Recap to obtain the appropriate format and then (c) modelling in Revit. Revit 2017 was used for this evaluation. The parameters of the cylinders (radius and length) are chosen based on the modeler's discretion. Non-cylindrical objects were not modelled in Revit since their extraction in EdgeWise is manual, thus a comparison with Revit is redundant.

3.5 Results

The time needed to model the above-mentioned object types is measured in the same operating system as stated above for pipeline, structural and electrical object types. The average modelling time per object for the most frequent object types is calculated. The manual modelling time of cylindrical objects is broken down to the two steps investigated above; shape extraction and semantic classification. Knowing the average number of objects of a specific type in a typical facility, we calculate the average modelling time for each object type and each modelling step where applicable. Figure 8 shows the modelling time/object in minutes and Fig. 9 the estimated total man-hours for modelling of the same object types in a typical industrial facility in hours.

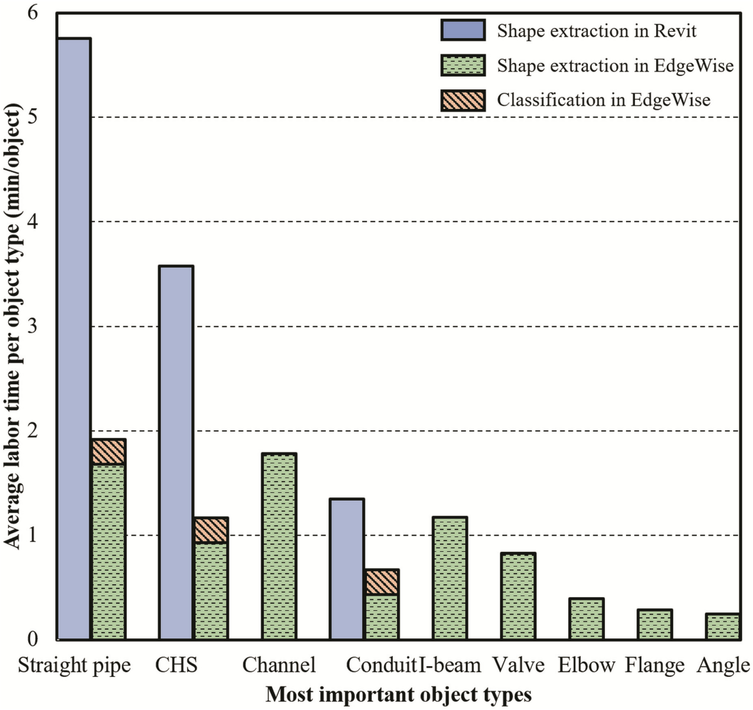


Fig. 8. Average modelling labor time per object (min/object) for the most frequent object types

Figure 8 shows that manual extraction of straight pipes in EdgeWise is the most time-intensive task compared to semantic classification for pipes and requires 1.68 min/straight pipe. Manual extraction of channels is also a laborious task compared to the manual extraction of all other object types, requiring 1.78 min/channel due to the complexity of their shape. Although some of the CHSs are automatically extracted, it is difficult to identify them manually, since they are usually pipe supports and handrails, which are significantly occluded. For instance, pipe supports are occluded due to pipelines that run on top of them. This is the reason for intensive modelling time (0.93 min/CHS). Semantic classification of cylinders is not a time-intensive step, requiring less than 0.5 min/cylinder on average.

Given the average frequencies of objects in each type obtained in [11] for a sample facility of 100,000 objects, Fig. 9 shows that pipes require the most modelling time on average (around 800 h) for this sample facility of 25,183 pipes. It is important to note that, although automated extraction of cylinders has been partially achieved by Edge-Wise Plant/MEP, modelling of pipelines takes still substantial amount of time. The cylindrical shape is the most frequent geometric shape, thus the modelers' effort to

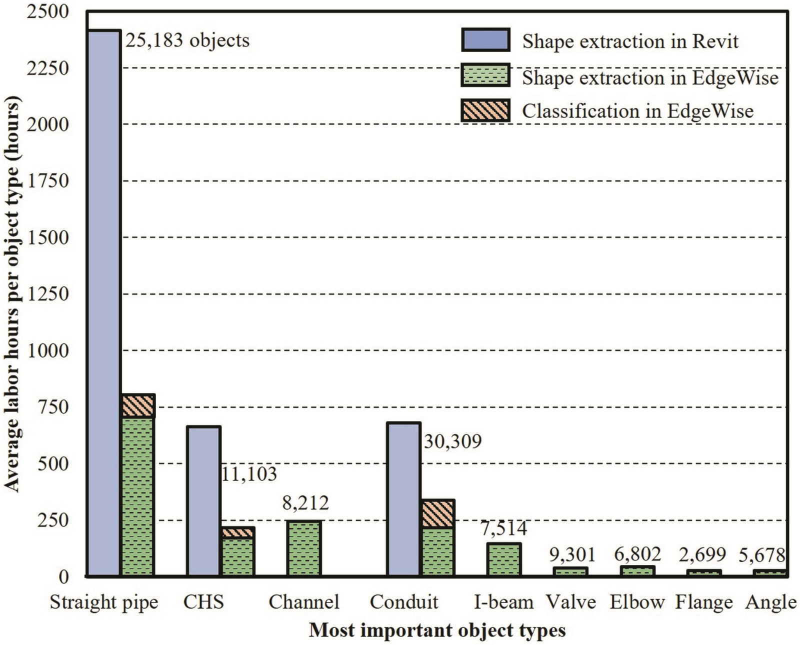


Fig. 9. Average modelling labor hours per object type for the most frequent objects of an example facility with shown numbers of objects

distinguish electrical conduit, CHSs, handrails and other cylindrical objects from straight pipes is significant.

Although electrical conduit is the most prevalent object type in industrial plants (24.3% in a typical plant, [11]), it takes less man-hours to model it compared to a straight pipe. This is attributed to the design of electrical conduit that places many cylinders closely to each other. This makes it easier for the modeler to identify them, thus the modelling time is reduced.

Flanges and elbows do not require substantial time (0.28 and 0.39 min/object respectively) as shown in Fig. 8, although the user manually adds them in the pipeline model. We observe that once the piping network is identified, the addition of fittings is a quick task that does not necessarily need to be automatically modelled. Angles require the least amount of time, being less than 0.25 min/angle, which is attributed to their simple geometry compared to I-beams or channels.

The total labor hours for manual modelling of an example industrial facility with 100,000 objects of the above categories are estimated to be 12 person-months. This finding is based on the following assumptions: (a) one trained modeler for all case studies, (b) the working hours are assumed to be 8 h/day, 5 days/week and (c) the operating system is as specified above. The same metric for cylinder extraction and classification is 8.5 person-months using EdgeWise as explained above. The confidence intervals for the average manual modelling time of pipes, conduit and CHSs are calculated since the selection of parameters depends on the modeler's discretion. Pipes were manually modelled in Revit in 5.8 ± 1 min with 99% confidence level. Conduit and CHSs were modelled in 1.3 ± 0.75 and 3.6 ± 0.4 min respectively. This means that the modelling time does not change substantially for any of these object categories.

The modeller was trained to model MEP and structural objects before starting the modelling task. 100 instances of each object type were modelled as a training exercise before the modeller started to perform this task. The theoretical example of 100,000 objects was chosen, given that the average total number of these object types is 191,991 as obtained from the case studies investigated in [11].

We observe that 64% of the man-hours needed for manual modelling of cylinders are saved by using the state-of-the-art software, EdgeWise, compared to conventional manual modelling platforms such as Revit. The results also show that 67% of manual modelling time is saved for pipe modelling in EdgeWise. The case study of 100,000 objects shows that 2,400 labor hours are saved when modelling cylinders in EdgeWise. This is crucial especially for these facilities, since the time required to take decisions for maintenance and refurbishment is limited due to continuous production flow.

4 Conclusions

The modelling time and shape extraction of the ten most frequent industrial object types are evaluated in the semi-automated, state-of-the-art software, EdgeWise. The results showed that cylindrical objects (straight pipes, electrical conduit and circular hollow sections) require 80% of the Total Modelling Hours (TMH) of the ten most frequent object types in EdgeWise.

The results of this paper show that current practice has achieved primitive shape extraction for straight pipes, elbows and conduit semi-automatically. However, semantic labelling of each object type is not performed in the state-of-the-art modelling packages. EdgeWise has substantially facilitated 3D modelling of industrial plants according to the findings discussed above. However, it has some limitations, which can be summarized as follows:

1. The modeler should identify the structural elements manually or define the location of an object roughly in the point cloud to fit it.
2. Detection of cylinders has only been partially solved, since cylinders are detected with 75% recall and 62% precision. The same metrics for pipes are 58% and 47% respectively.
3. EdgeWise and all other 3D modelling software platforms do not enrich the 3D geometric primitives with semantic labels and topological relationships. Engineers are required to manually implement the semantic labels of the components of the 3D model.
4. Data inconsistency between different software platforms impedes modelers from exchanging data between different AI-BIM platforms. These software packages are not designed to provide a final output in an open and generic schema.

The contribution of this paper is the measurement of the performance of state-of-the-art software and more specifically EdgeWise. This uncovered (a) the substantial performance of this software in detecting cylinders, (b) the inability of this software to (i) further classify cylinders into electrical conduit, pipes or CHSs and (ii) detect and further classify I-beams, channels, elbows, flanges, valves and angles in spite of their high frequency in an industrial facility.

Direct implications of modelling the most frequent industrial objects of [11] are assessed based on modelling time. The results of the evaluation of EdgeWise showed that semi-automatically modelling cylinders will reduce man-hours needed for modelling those by 64%. This can have a direct impact for industrial facility managers, since every hour of as-is modelling time is crucial for the operation of the plant in unprecedented circumstances (failures of critical objects, retrofitting operations and plant expansion).

Indirect implications of prioritizing object types are reductions of the modelling cost, since man-hours of modelers will be reduced. Although there is no way to calculate the exact cost of overestimated severity of industrial inspections and maintenance, it is reasonable to predict that maintenance of industrial plants will be substantially facilitated once AI-BIMs are easy to develop and the costs do not counteract the benefits of their creation. Poor maintenance of these assets does not always affect the asset's territory but also impacts nearby regions and puts lives of the public living close by at serious risk.

The presented research has room for improvement and some limitations of this study can direct future research. This study focused on the industrial objects that are important to model, however methods on how to automatically model those were not investigated. Current research efforts were compared with commercial tools like EdgeWise showing no significant advances in terms of automated detection of cylinders. Future work involves implementation of automated machine learning algorithms for all the most frequent object types to minimize the modelling time. Application of these algorithms for hundreds of classes of different objects is a difficult multi-classification problem, that will be substantially benefited from the results of this exploratory research for the most frequent objects to model in these complex environments. Overall, a training library of the object classes that are critical for industrial facility operations, frequent in industrial environments and laborious to model can be established to assist further research aimed at automated detection of these classes. Application of the findings of this paper will guide researchers on investigating methods for automatically modelling these objects.

Acknowledgements. We thank our colleague Graham Miatt, Senior Software Engineer in LFM Software from AVEVA Group Plc., who provided insight and expertise that greatly assisted this research. We also express our gratitude to Otto Weiberth, Managing Director of ClearEdge3D, Inc., who provided EdgeWise Suite for evaluation. The research leading to these results has received funding from the Engineering and Physical Sciences Research Council (EPSRC). AVEVA Group Plc. and BP International Centre for Business and Technology (ICBT) partially sponsor this research under grant agreements RG83104 and RG90532 respectively. We gratefully acknowledge the collaboration of all academic and industrial project-partners. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the institutes mentioned above.

References

1. Volk, R., Stengel, J., Schultmann, F.: Building Information Modeling (BIM) for existing buildings - literature review and future needs. Autom. Constr. **38**, 109–127 (2014)
2. Tornincasa, S., Di Monaco, F.: The future and the evolution of CAD. In: International Research Conference on Trends Development Machine Association Technology, vol. 18 (2010)
3. Cabinet Office H: Government Construction Strategy. Construction 96:43, vol. 19 (2011)
4. Rabbani, T., van den Heuvel, F.A., Vosselman, G.: Segmentation of point clouds using smoothness constraint. In: Proceedings of the Image Engineering and Vision Metrology, vol. 36, pp. 248–253. International Society for Photogrammetry and Remote Sensing (2006)
5. Son, H., Kim, C., Kim, C.: Knowledge-based approach for 3D reconstruction of as-built industrial plant models from laser-scan data. In: International Symposium on Automation and Robotics in Construction, vol. 756, pp. 885–893 (2013)
6. Kawashima, K., Kanai, S., Date, H.: As-built modeling of piping system from terrestrial laser-scanned point clouds using normal-based region growing. J. Comput. Des. Eng. **1**, 13–26 (2014). https://doi.org/10.7315/jcde.2014.002
7. Veldhuis, H., Vosselman, G.: The 3D reconstruction of straight and curved pipes using digital line photogrammetry. ISPRS J. Photogramm. Remote Sens. **53**, 6–16 (1998). https://doi.org/10.1016/s0924-2716(97)00031-2
8. Sanders, F.H.: 3D laser scanning helps Chevron revamp platform. Oil Gas J. **99**, 92–98 (2001)
9. Fumarola, M., Poelman, R.: Generating virtual environments of real world facilities: discussing four different approaches. Autom. Constr. **20**, 263–269 (2011)
10. Hullo, J.-F., Thibault, G., Boucheny, C., et al.: Multi-sensor as-built models of complex industrial architectures. Remote Sens. **7**, 16339–16362 (2015). https://doi.org/10.3390/rs71215827
11. Agapaki, E., Brilakis, I.: Prioritising object types of industrial facilities to reduce As-Is modelling time. In: Proceedings of the 33rd Annual ARCOM Conference, Cambridge, U.K., 4–6 September 2017, pp. 402–411 (2017)
12. Douglas, J.M.: Conceptual Design of Chemical Processes. McGraw-Hill, New York (1988)
13. FARO: FARO (2017). https://www.faro.com/en-in/products/construction-bim-cim/faro-pointsense/features/
14. Ahmed, M.F., Haas, C.T., Haas, R.: Automatic detection of cylindrical objects in built facilities. J Comput. Civ. Eng. **28**, 4014009 (2014). https://doi.org/10.1061/(asce)cp.1943-5487.0000329
15. Patil, A.K., Holi, P., Lee, S.K., Chai, Y.H.: An adaptive approach for the reconstruction and modeling of as-built 3D pipelines from point clouds. Autom. Constr. **75**, 65–78 (2017). https://doi.org/10.1016/j.autcon.2016.12.002
16. Son, H., Kim, C., Kim, C.: Knowledge-based approach for 3D reconstruction of as-built industrial plant models from laser-scan data. In: ISARC, vol. 756, pp. 885–893 (2013)
17. Perez-Gallardo, Y., Cuadrado, J.L.L., Crespo, Á.G., de Jesús, C.G.: GEODIM: a semantic model-based system for 3D recognition of industrial scenes. In: Intelligent Systems Reference Library, pp. 137–159 (2017)
18. ClearEdge: ClearEdge (2017). http://www.screencast.com/users/ClearEdge3D
19. Hackel, T., Wegner, J.D., Schindler, K.: Contour detection in unstructured 3D point clouds. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1610–1618 (2016)
20. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. **2**, 37–63 (2011). https://doi.org/10.1016/j.isprsjprs.2017.05.012

Pavement Defects Detection and Classification Using Smartphone-Based Vibration and Video Signals

Symeon E. Christodoulou(✉), Georgios M. Hadjidemetriou, and Charalambos Kyriakou

Department of Civil and Environmental Engineering, University of Cyprus, Nicosia, Cyprus
schristo@ucy.ac.cy

Abstract. Presented herein is a big-data driven methodology for the detection of roadway anomalies, utilizing smartphone-based data and image signal streams. The methodology uses a vibration-based method and artificial intelligence for the detection of vibration-inducing anomalies, and a vision-based method with entropic texture segmentation filters and support vector machine (SVM) classification for the detection of patch defects on roadway pavements. The presented system pre-processes video streams for the identification of video frames of changes in image-entropy values, isolates these frames and performs texture segmentation to identify pixel areas of significant changes in entropy values, and then classifies and quantifies these areas using SVMs. The developed SVM is trained and tested by feature vectors generated from the image histogram and two texture descriptors of non-overlapped square blocks, which constitute images that includes "patch" and "no-patch" areas. The outcome is composed of block-based and image-based classification, as well as of measurements of the patch area.

Keywords: Pavement condition evaluation · Video signal · Acoustic signal Smartphones

1 Introduction

In recent years several national and transnational roadway management programs, such as the USA's Long-Term Pavement Performance (LTPP) and the EU's Trans-European Transport Network (TEN-T) programs, have been put to action in an effort to improve on the condition of transport networks and to mitigate the effects of time and of heavy usage on these networks. In fact, several regional and international studies estimate the annual potential impacts of changes in roadway maintenance expenditures (as these impacts relate to vehicle operating costs, safety, the environment, and the wider economy) at billions of dollars worldwide (Chatti and Zaabar 2012; National Economic Council 2014; Gleave et al. 2014). It is thus becoming increasingly important that automated roadway pavement condition assessment technologies are employed, so that sustainable and efficient roadway network management systems are developed. A key

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 125–138, 2018.
https://doi.org/10.1007/978-3-319-91635-4_7

function of such technologies should be the automatic roadway defects detection and classification.

Pavement surface condition, determined by the anomalies in the pavement surface that have an effect on the ride quality of a vehicle, is one of the most significant indicators for road quality. Not only can pavement surface anomalies damage vehicles and cause unpleasant driving but they are also the source of traffic accidents, injuries and/or fatalities. Further, the condition of pavement surfaces is in constant flux over time, as pavements deteriorate from causes related to location, materials used, traffic, weather, etc.

Currently, even though the continuous monitoring of pavement surface defects would improve the ride quality and the safety of travelers, pavement monitoring agencies typically monitor assess pavement quality approximately once per year because current methods are expensive and laborious. The rise, though, of low-cost technologies could provide a viable solution to the aforementioned problem in the form of distributed mobile sensing by use of connected vehicles and smartphones.

In current years, smartphone technology has gained significant attention within the infrastructure, transportation, and automotive industries. Smartphones can be exploited to collect vehicle sensor data by use of their built-in sensors (such as accelerometer, gyroscope and GPS sensors), and/or linked to on-board diagnostic devices (OBD) to collect data from the host vehicle on how the vehicle performs whilst in movement. This combination of hardware and software mechanisms can enable the real-time monitoring of, among others, GPS latitude and longitude, forward and lateral acceleration, vehicle roll and pitch.

The goal for roadway anomaly detection by use of smartphone technology, OBD devices and vehicles is set in parallel with the premise that such technology can be applied in GIS-based pavement management systems (PMS). An adequate number of vehicles collecting this crowd-sourced data can be used in order to generate georeferenced events at points where vehicles encounter pavement surface anomalies within a roadway network. Even though multiple vehicles may probably provide conflicting data with regard to pavement surface conditions, the total effect and the joint ‘knowledge provided by participatory sensing’ which is inherent in the collected data do provide an accurate model of the pavement surface in relation to how an average user experiences the roadway condition.

With regard to the volume and complexity of data processed, “big data” techniques may provide viable solutions to the challenges posed by the problem. The term “big data” does not necessarily refer to the size of the datasets being processed (even though in this case the sensor and image-related data is voluminous), but it often refers simply to the use of predictive and other advanced data analytics methods that extract value from data.

The paper presents both a data-driven vibration-based method for the detection and classification of pavement anomalies by use of low-cost (smartphone) technology, and a vision-based method for the enhancement of the detection process. The vibration-based component makes use of artificial neural networks (ANN) for data-mining and of timeseries analysis of vibration signals to detect vibration-inducing roadway anomalies, whilst the vision-based method uses video data, image segmentation via entropy texture filters, and object classification via support vector machines (SVM) to detect roadway anomalies.

Further to this brief introduction the paper includes a literature review on the state of knowledge in automated roadway anomaly detection, followed by brief discussions of the methods used in the analysis: data-mining, ANN, texture segmentation, entropy filters, and SVMs. The section on methodology setup presents the developed data collection system and methods, while the results and discussion section discusses the processes and tools used to detect and classify pavement anomalies. The main characteristics of a case-study pothole-detection implementation are then presented, a patch detection analysis is performed for a case-study roadway and finally the results and efficiency of the proposed methods are discussed.

2 Brief Literature Review

The automated detection of roadway pavement anomalies by use of low-cost technologies has been the focus of several research efforts in the past decade, with these efforts generally classified in two categories: vibration-based and vision-based methods. A brief summary of some of these approaches and of their findings is listed below.

2.1 Vibration-Based Methods

Vittorio et al. (2014) proposed a system based on a simple application for smartphones that uses a GPS receiver and a three axis accelerometer to collect acceleration data due to vehicles motion on road anomalies. The high-energy events (anomalies) are identified by monitoring and measuring the vertical acceleration impulse. Seraj et al. (2014) proposed a system that detects road anomalies using mobile phones equipped with inertial accelerometers and gyroscopes sensors. Alessandroni et al. (2014) proposed a system which combined a custom mobile application and a georeferenced database system. The roughness score was calculated and stored into a back-end geographic information system for visualizing road conditions. Mohamed et al. (2015) in order to avoid false-positive signals when there was a sudden change in motion acceleration suggested the gyroscope around gravity rotation as the main indicator for road anomalies, in addition to the accelerometer sensor. Jang et al. (2016) proposed an automated method to obtain up-to-date information about potholes by using a mobile data collection kit, mounted on vehicles. In each mobile data collection kit, a triaxial accelerometer and global positioning system sensor collect data for the detection of street defects. At a back-end server, a street defect algorithm which relies on a supervised machine learning technique and a trajectory clustering algorithm enhances the performance of the proposed monitoring system. The above systems, despite hardware differences in terms of GPS accuracy and accelerometer sampling rate and noise, they show that pothole detection is possible.

Bridgelall (2015) developed theoretical precision bounds for a ride index called the road impact factor and demonstrated its relationship with vehicle suspension parameter variances. The 2014 Mercedes-Benz S-Class used a Light-Detection-and-Ranging (lidar) scanner to estimate pavement surface roughness as a part of an active suspension system.

Lately, the Jaguar Land Rover automaker make known that is testing a new connected vehicle technology which permits a vehicle to spot hazardous potholes in the roadway and then distribute this data in real time with other Jaquar Land Rover vehicles (O'Donnell 2015). Kyriakou et al. (2016, 2017) explored the use of data collected by sensors from smartphones and vehicles utilizing automobiles' OBD-II devices while vehicles were in movement, for the detection and classification of pavement surfaces anomalies. The proposed system architecture was complimented with artificial neural network techniques for classifying detected roadway anomalies. The proposed system was trained, validated and tested against three types of common roadway anomalies exhibiting above 90% accuracy rate.

2.2 Vision-Based Methods

In the work by Nejad and Zakeri (2011) an automated imaging system was described for distress detection in asphalt pavements. The work focused on comparing the discriminating power of several multi-resolution texture analysis techniques using wavelet, ridgelet and curvelet-based texture descriptors, and concluded that curvelet-based signatures outperform all other multi-resolution techniques for pothole distress, (yielding accuracy rates of 97.9%), while ridgelet-based signatures outperform all other multi-resolution techniques for cracking distress (accuracy rates of 93.6%–96.4%).

A computer-vision approach was also the subject of the work by Koch and Brilakis (2011), who proposed a method for automated pothole detection by which an image was first segmented into defect and non-defect regions using histogram shape-based thresholding, and then the texture inside a potential defect shape was extracted and compared with the texture of the surrounding non-defect pavement in order to determine if the region of interest represents an actual pothole. The aforementioned camera-based pothole-detection method was subsequently extended by Koch et al. (2013) for assessing the severity of potholes, by incrementally updating a representative texture template for intact pavement regions and using a vision tracker to reduce the computational effort. Related was also the work by Jog et al. (2012) who used vision-based data for both 2D recognition and for 3D reconstruction, based on visual and spatial characteristics of potholes, and measured properties were used to assess the severity of potholes.

Citing limitations of camera-based methods, Yu and Salary (2011) proposed the use of laser imaging and described a method by which regions in captured images corresponding to potholes are represented by a matrix of square tiles and the estimated shape of the pothole is determined. The vertical, horizontal distress measures, the total number of distress tiles and the depth index information are calculated providing input to a three-layer feed-forward neural network for pothole severity and crack type classification. A vision approach was also employed by Murthy and Varaprasad (2014) who used images obtained from a camera mounted on top of a vehicle and custom MATLAB code to detect potholes. In the work by Ryu et al. (2015) a pothole detection method was proposed using various features in two-dimensional images. The proposed method first uses a histogram and the closing operation of a morphology filter to extract dark regions for pothole detection, and then candidate regions of a pothole are extracted with the use of features such as size and compactness. Finally, a decision is made on whether candidate regions are potholes with a comparison of pothole and background features. Radopoulou and Brilakis (2015) presented an application of the Semantic Texton Forests (STF) algorithm for automatically detecting patches, potholes and three types of cracks in video frames captured by a common parking camera, reporting over 70% accuracy in all of the tests performed, and over 75% precision for most of the defects. Subsequently, Radopoulou et al. (2016) utilized video data collected from a car's parking camera to detect defects in frames and classified detected defects according to their type and severity. The researchers reported that the initial identification of frames including defects produced an accuracy of 96% and approximately 97% precision. A vision-based approach was, finally, employed by Li et al. (2016) who proposed a method to integrate the processing of two-dimensional images and of ground penetrating radar (GPR) data for pothole detection. The images and GPR scans are first pre-processed and a pothole detector designed by investigating the patterns of GPR signals, and then the position and dimension of the detected potholes is estimated from GPR data and mapped to the image to enable a localized shape segmentation. The researcher reported a precision, recall, and accuracy of 94.7%, 90% and 88%, respectively.

3 Methodology

3.1 Anomaly Detection Using ANNs and Timeseries Analysis of Vibration Signals

The described research work focusses on three types of common pavement surface anomalies (transverse depressions/patches, Fig. 1a; longitudinal depressions/patches, Fig. 1b; and potholes or manholes, Fig. 1c). Data on these types of pavement surface anomalies is collected in-situ by use of a car equipped with a smartphone (mounted on the car's windshield, having a 16MP camera, and video-recording at 640 × 480 pixel resolution) with its sensors turned on, and with an OBD-II reader attached to it. The smartphone was also fitted with an android application for recording (and exporting) sensor readings of taken data. Further for visually verifying the existence of a pavement surface anomaly (as detected by the sensors data), the smartphone had also its video camera active for recording the routes travelled.

The sensor data, collected at intervals of 0.1 s, included a total of 14 uni-dimensional (e.g. X, Y, Z accelerations, speed, etc.) and two-dimensional indicators (e.g. the smartphone's roll and pitch values). The smartphone was also fitted with the *DashCommand* application for recording sensor readings of taken data. Vehicle system data is transmitted through the OBD-II reader to the smartphone device and then transferred for to a back-end server for processing and storing. At the back-end server, defect detection algorithms based on artificial neural network techniques, robust regression analysis, various algorithms and bagged trees classification model enhances the performance of the proposed monitoring system, by integrating data collected from multiple sensors and deducing knowledge from these participatory sensors. Mathematically, the proposed method is based on rigid-body dynamics and in particular the roll, pitch and yaw rotations about the object's XYZ axis (Kyriakou et al. 2017). In essence, the roll metric refers to a car's acceleration variation between its left and

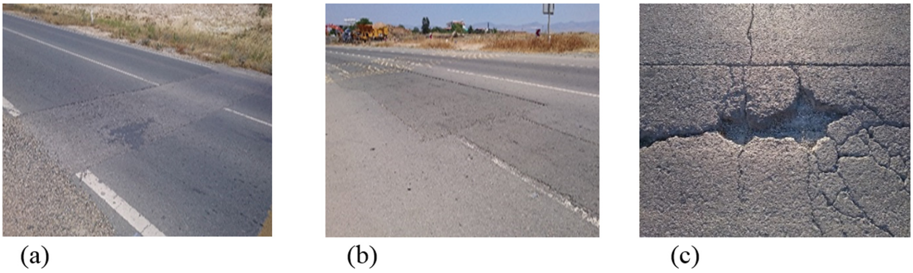


Fig. 1. Pavement surface anomaly types examined for detection and classification: (a) transverse defect/anomaly; (b) longitudinal defect/anomaly; (c) potholes/manholes.

right front wheels, while the pitch metric refers to a car's acceleration variation between its front and rear wheels. Concurrently, the roll and pitch values define the way in which the host car is off balance (sideways and front/back).

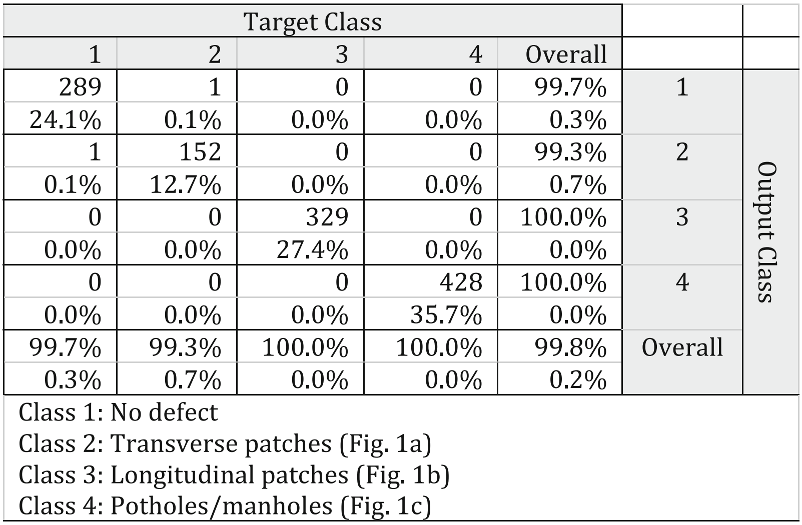
The datasets are fed into an artificial neural network (ANN) consisting of 4 inputs (Forward Acceleration, Lateral Acceleration, Vehicle Pitch, Vehicle Roll), 10 hidden neurons and 4 outputs (Class Type 1, Class Type 2, Class Type 3, Class Type 4) representing the roadway anomalies listed in Fig. 2. The ANN outputs are binary in nature ('0' for no defect, '1' for defect) and they are used to classify data readings into classes of roadway anomalies. The ANN is first trained for each case of roadway anomaly (as given by Fig. 1), and then trained with all three roadway anomalies in tandem (Class Type 1, Class Type 2, Class Type 3, Class Type 4).

The ANN is first trained with each case of roadway anomaly (as given by Fig. 1), and then trained with all three roadway anomalies in tandem (ClassType_0, ClassType_1, ClassType_2, ClassType_3). For each case, 70% of the data is used for training, 15% for testing and 15% for validating the ANN. The results of the ANN-based pattern recognition and pavement anomaly classification are as shown in the confusion matrix of Fig. 2. In essence, the ANN classifier detects and accurately categorizes the three roadway anomalies (target classes '2', '3' and '4') while also distinguishing the 'no defect' condition (target class '1'), thus separating normal and abnormal roadway pavement conditions.

The vibration-based method, though, fails to efficiently cover the entire roadway and, more importantly, fails to detect non-vibration-inducing pavement anomalies. Hence, the need for a vision-based method to complement the aforementioned vibration-based approach.

3.2 Anomaly Detection Using Entropic-Filter Image Segmentation

The proposed vision-based approach makes use of image texture segmentation with entropy texture filters, and has been implemented on MATLAB's computer vision toolbox. Entropy is a statistical measure of randomness, and an entropy filter can characterize the texture of an image by providing information about the local variability of the intensity values of pixels in an image. For example, in areas with smooth texture, the range of values in the neighborhood around a pixel will be a small value; in areas of rough



Target Class						
1	2	3	4	Overall		
289	1	0	0	99.7%	1	Output Class
24.1%	0.1%	0.0%	0.0%	0.3%		
1	152	0	0	99.3%	2	
0.1%	12.7%	0.0%	0.0%	0.7%		
0	0	329	0	100.0%	3	
0.0%	0.0%	27.4%	0.0%	0.0%		
0	0	0	428	100.0%	4	
0.0%	0.0%	0.0%	35.7%	0.0%		
99.7%	99.3%	100.0%	100.0%	99.8%	Overall	
0.3%	0.7%	0.0%	0.0%	0.2%		

Class 1: No defect
Class 2: Transverse patches (Fig. 1a)
Class 3: Longitudinal patches (Fig. 1b)
Class 4: Potholes/manholes (Fig. 1c)

Fig. 2. ANN confusion matrix.

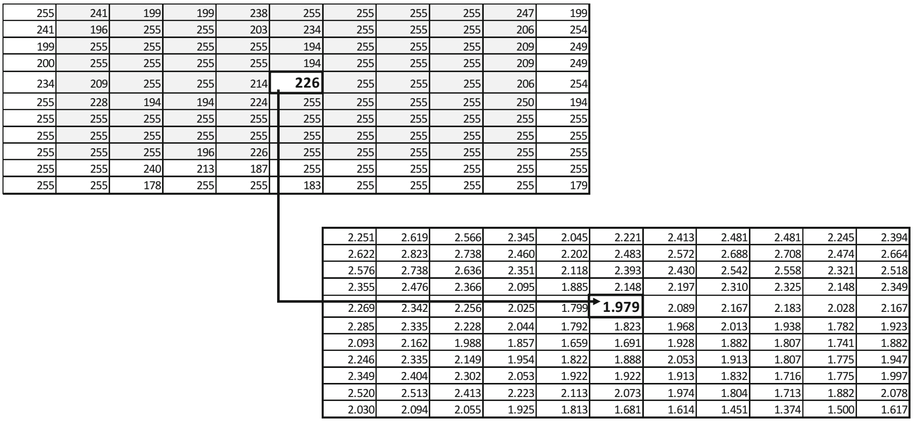
texture, the range will be larger. Similarly, calculating the standard deviation of pixels in a neighborhood can indicate the degree of variability of pixel values in that region.

The entropy (E) of a grayscale image (I) is defined as $E = -sum[p.*log2(p)]$, where p contains the histogram counts of the intensity image. By default, entropy uses two bins for logical arrays and 256 bins for uint8, uint16, or double arrays. The entropy filter ($J = entropyfilt(I)$) of a grayscale image returns the array (J), where each output pixel contains the entropy value of the 9-by-9 neighborhood around the corresponding pixel in the input image I (Fig. 3). Thus, the entropy filter creates a texture image. For pixels on the borders of I, *entropyfilt* uses symmetric padding, where the values of padding pixels are a mirror reflection of the border pixels in I.

The steps used in the proposed entropy texture segmentation approach are as listed below, with Fig. 4 serving as a reference for the resulting image at each analysis step:

1. Read into MATLAB a video of the roadway pavement to be analysed (function '*VideoReader*').
2. For each video frame, convert it to grayscale image (Fig. 4a) and calculate the overall image entropy (functions '*read*', '*rgb2gray*' and '*entropy*').
3. If the computed image entropy deviates from the running average, then presume that the image contains a pavement anomaly (manifested in the image as texture anomaly) and isolate it for further analysis.

- Create a texture image (Fig. 4b).
- Threshold the image to segment the textures (a threshold value of 0.8 is used as default value, for it is roughly the intensity value of pixels along the boundary between the textures). A function is also used to smooth the edges and to close any open holes in objects.



255	241	199	199	238	255	255	255	255	247	199
241	196	255	255	203	234	255	255	255	206	254
199	255	255	255	255	194	255	255	255	209	249
200	255	255	255	255	194	255	255	255	209	249
234	209	255	255	214	**226**	255	255	255	206	254
255	228	194	194	224	255	255	255	255	250	194
255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255
255	255	255	196	226	255	255	255	255	255	255
255	255	240	213	187	255	255	255	255	255	255
255	255	178	255	255	183	255	255	255	255	179

2.251	2.619	2.566	2.345	2.045	2.221	2.413	2.481	2.481	2.245	2.394
2.622	2.823	2.738	2.460	2.202	2.483	2.572	2.688	2.708	2.474	2.664
2.576	2.738	2.636	2.351	2.118	2.393	2.430	2.542	2.558	2.321	2.518
2.355	2.476	2.366	2.095	1.885	2.148	2.197	2.310	2.325	2.148	2.349
2.269	2.342	2.256	2.025	1.799	**1.979**	2.089	2.167	2.183	2.028	2.167
2.285	2.335	2.228	2.044	1.792	1.823	1.968	2.013	1.938	1.782	1.923
2.093	2.162	1.988	1.857	1.659	1.691	1.928	1.882	1.807	1.741	1.882
2.246	2.335	2.149	1.954	1.822	1.888	2.053	1.913	1.807	1.775	1.947
2.349	2.404	2.302	2.053	1.922	1.922	1.913	1.832	1.716	1.775	1.997
2.520	2.513	2.413	2.223	2.113	2.073	1.974	1.804	1.713	1.882	2.078
2.030	2.094	2.055	1.925	1.813	1.681	1.614	1.451	1.374	1.500	1.617

Fig. 3. Sample *entropyfilt()* calculations

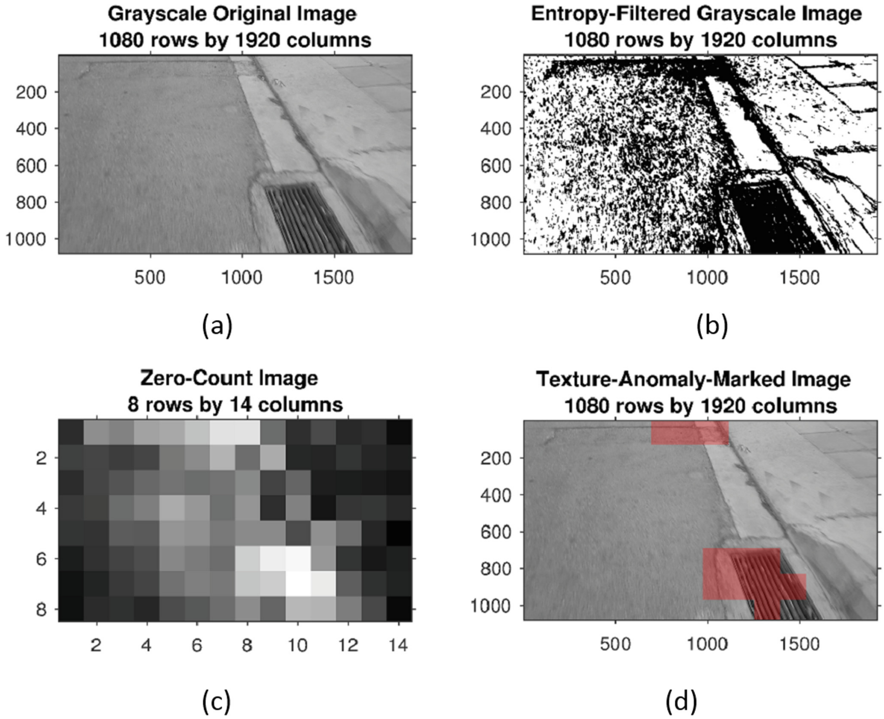


Fig. 4. Sample texture segmentation and anomaly detection using entropy filters.

- Partition the entropy image (Fig. 4b) into a grid (in this case from 1080 × 1920 pixels to 18 × 32 pixels), and count the proportion of black-to-white pixels in each grid cell. Threshold this ratio (say at 80%) and output the resulting image (Fig. 4c). White regions indicate texture anomalies.
- Rescale the threshold texture image (Fig. 4c) back to the original image dimensions and display the segmentation results, marking the corresponding image areas as pavement anomalies (Fig. 4d).

3.3 Patch Detection and Measurement Using Support Vector Machines (SVM)

The texture segmentation approach has been complimented with support vector machines (SVM) and applied to patch detection and measurement. SVMs are supervised machine learning models, which identify patterns after taking labelled training data. An SVM is divided into two main phases (training and testing), and it can be used efficiently for two-group classification problems such as the presented research study ("patch" vs. "no-patch" classes).

The key steps of the presented algorithm (Hadjidemetriou et al. 2016, 2017) are presented in Fig. 5 for both the training and testing stages, and the input consists of the pavement surface images extracted from the roadway videos. The SVM training phase begins with transforming collected pavement video frames into grayscale images. Thus, every pixel has a grayscale value in the range from 0 (which characterizes black) and 255 (which describes white). The presented system uses only one SVM, which is trained by labelled data (ground truth) and feature vectors.

The ground truth is entered in the algorithm given data regarding the pixels of each frame which are part of a pavement patch. Each feature vector is generated, and subsequently the SVM is trained, extracting information from non-overlapped areas within the frame, whose size is 20 × 20 pixels in width and height. The selection of blocks size is based on usual image resolutions, whose dimensions are multiples of 20 (e.g. 640 × 480), so that blocks would cover the whole image. A number of block sizes, which fulfil this criterion (e.g. 10 × 10), have been tested and the use of a trial and error technique designates our final district size (20 × 20). One should also note that blocks which are comprised of weighty proportions of both patch and non-patch areas (i.e. the patch area is more than 5% and less than 95% of the block) are not used to facilitate the training of the SVM and consequently its ability to distinguish "patch" from "no-patch" areas. Every feature vector, corresponding to a square block, is generated by the local intensity histogram and two texture descriptors, named two-dimensional Discrete Cosine Transform (DCT) and Gray-Level Co-occurrence Matrix (GLCM). DCT, which can be used efficiently for purposes of pattern recognition purposes, expresses a finite amount of data points in respect of a weighted sum of cosine functions oscillating at diverse frequencies. GLCM is a statistical system that examines the spatial relationship of pixels; while its functions are able to designate the texture of a picture by creating a matrix, which contains the estimated frequencies of the occurrence of pixel pairs with definite values and in a specific spatial relationship. The presented method extracts data from this matrix to calculate and then uses the statistical measures of contrast, correlation, energy and homogeneity.

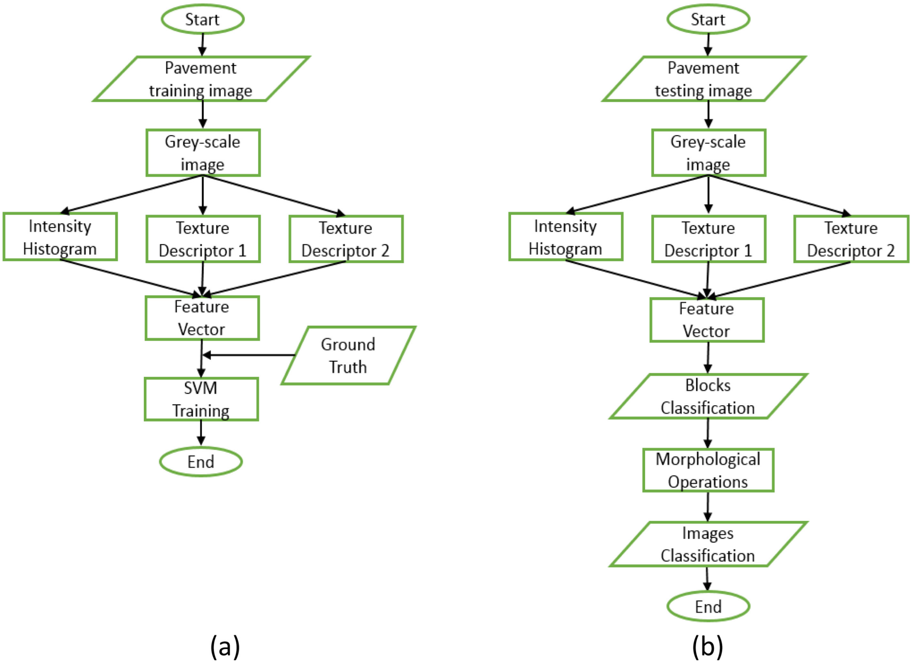


Fig. 5. Training and testing stages of the proposed patch detection algorithm.

The SVM training stage is followed by a testing phase (Fig. 5b). Its flow is similar with the SVM training, starting with transforming RGB pavement frames into grayscale images and dividing them into square blocks of 20 × 20 pixels. A feature vector is formed by the local intensity histogram and the two texture descriptors for each square block. The flowchart continues with the feature vector used by the SVM to classify each block of the testing pavement picture in "patch" (1) or "no-patch" (0) categories. Figure 6 depicts the way patch areas are identified by the algorithm; where yellow-colored blocks represent the "patch" class and blue-colored cells correspond to the "no patch" group.

Further, the morphological operation of closing is applied, to fill and eliminate blocks which are classified differently than their surrounded blocks, by changing their label from 0 to 1 and vice versa (Fig. 7). Finally, a trial and error technique is used to define the number of connected "patch" blocks (50) which indicate the presence of a patch in an image. Figure 7 presents an example of a frame correctly classified as "including patches" as it has more than 50 connected blocks; and a second example, truly identified as "not including patches", even if it has 46 of False Positive connected blocks. In case a patch has just appeared in the video view (thus covering a limited proportion of the frame) and the identified connected blocks are less than 50, then it will not be detected by the algorithm. However, the next extracted image from the video will include a greater percentage of the patch, providing the opportunity to be detected. Consequently, the algorithm, after the blocks classification, discriminates

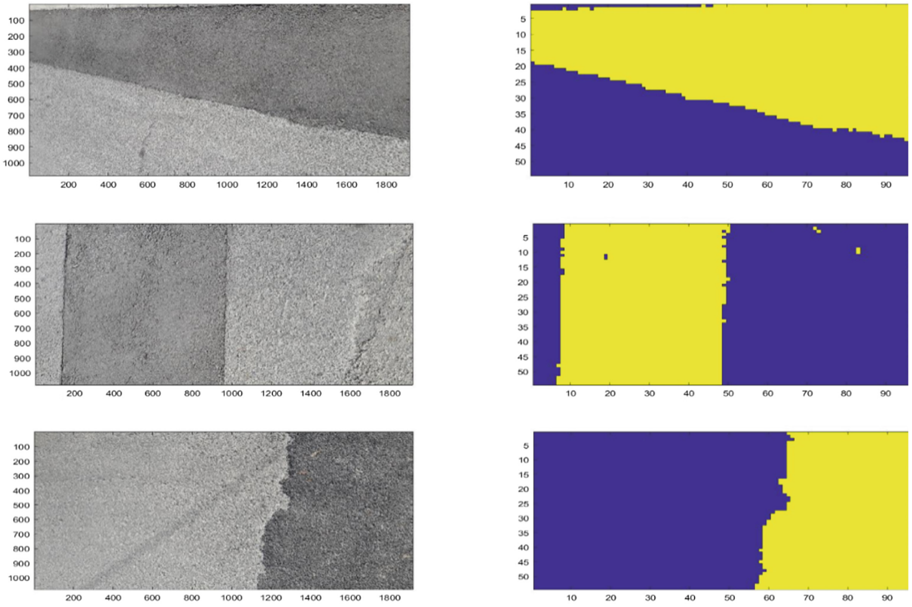


Fig. 6. Examples of processed images by the proposed algorithm. (Color figure online)

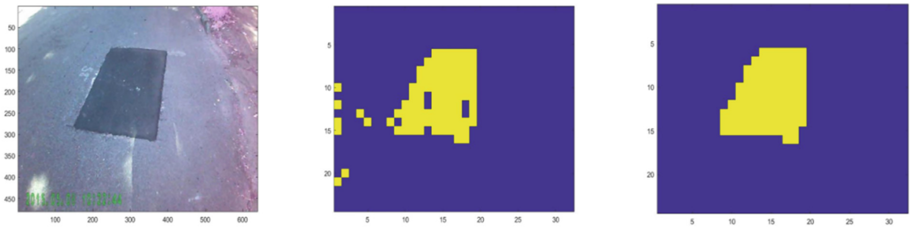


Fig. 7. Application of the morphological operation.

between the images which include parts of patches and the frames which do not contain any patches parts (image classification). At this point, the difference between presence and detection should be clarified. The former answers with a “yes” or a “no” the question of whether an examined object occurs in an image, while the latter provides information regarding the place of the object in the image. Despite the restriction of a range of algorithms to only identifying the presence of a distress, while classifying images between damaged and undamaged pavement, the proposed algorithm achieves both the presence and detection of patches.

The method has been successfully field-tested on case-study roadways, using either a generic dash cam or a smartphone camera, with the accuracy, precision and recall rates shown in Table 1.

Table 1. The performance of blocks and images classification - (a) generic dash camera; (b) smartphone camera.

Block classification		Images classification		Blocks classification		Images classification	
(Generic dash camera)				(Smartphone camera)			
Accuracy	82.9%	Accuracy	82.5%	Accuracy	80.5%	Accuracy	80.0%
Precision	65.6%	Precision	77.8%	Precision	63.8%	Precision	75.4%
Recall	92.0%	Recall	91.0%	Recall	89.4%	Recall	89.0%

Further to the obtained accuracy levels, the presented method is characterized by some strong advantages such as the identification of multiple patches in a single image or the detection of proportions of patches when their entire area is not included in the image.

4 Conclusions

The paper presented a vibration-based and a vision-based method, working in tandem, for the detection and classification of roadway anomalies by use of “big-data” tools and low-cost smartphone technology. The popularity of smartphone technology in vehicles and the advancement of “big-data” technologies provide an opportunity to efficiently collect vehicle data and process it by use of connected and distributed systems.

Even though vehicle data is not likely to directly provide traditional assessment metrics (such as IRI and PCI), new metrics might supplement and eventually supplant traditional metrics. The applied methodology is instantly available, low-cost and precise, and can be utilized in crowd-sourced applications leading to roadway assessment and pavement management systems.

The described two approaches (vibration-based and vision-based) are complimentary in nature and advantages, and in tandem provide a low-cost worthy alternative to expensive methods. The vibration-based approach does not require heavy computational loads and extensive bandwidth for real-time processing, but it fails to efficiently cover the entire roadway. More importantly, it fails to detect non-vibration-inducing pavement anomalies. The vision-based approach, on the other hand, covers larger areas and it is not limited to tire-hit and vibration-inducing roadway anomalies.

Ongoing research work on both vibration-based and vision-based methods extends the applicability of the proposed methods to include additional roadway anomaly defects, such as cracks and raveling, and to increase the methods’ detection accuracy.

References

Alessandroni, G., Klopfenstein, L., Delpriori, S., Dromedari, M., Luchetti, G., Paolini, B., Seraghiti, A., Lattanzi, E., Freschi, V., Carini, A.: SmartRoadSense: collaborative road surface condition monitoring. In: The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, UBICOMM, Rome, Italy, pp. 24–28 (2014)

Bridgelall, R.: Precision bounds of pavement deterioration forecasts from connected vehicles. J. Infrastruct. Syst. **21**(1), 1–7 (2015)

Chatti, K., Zaabar, I.: Estimating the effects of pavement condition on vehicle operating costs. Report 720, Transportation Research Board (2012)

Gleave, S.D., Frisoni, R., Dionori, F., Casullo, L., Vollath, C., Devenish, L., Spano, F., Sawicki, T., Carl, S., Lidia, R., Neri, J., Silaghi, R., Stanghellini, A.: EU road surfaces: Economic and safety impact of the lack of regular road maintenance. European Parliament - Directorate General for Internal Policies, Policy Department B: Structural and Cohesion Policies, Transport and Tourism (2014)

Hadjidemetriou, G.M., Christodoulou, S.E., Vela, P.A.: Automated detection of pavement patches utilizing support vector machine classification. In: 18th Mediterranean Electrotechnical Conference (MELECON), Limassol, Cyprus, pp. 1–5. IEEE (2016)

Hadjidemetriou, G.M., Vela, P.A., Christodoulou, S.E.: Automated pavement patch detection and quantification using support vector machines. J. Comput. Civ. Eng. **32**(1), 04017073 (2017)

Jang, J., Yang, Y., Smyth, A., Cavalcanti, D., Kumar, R.: Framework of data acquisition and integration for the detection of pavement distress via multiple vehicles. J. Comput. Civ. Eng. **31**, 04016052 (2016)

Koch, C., Brilakis, I.: Pothole detection in asphalt pavement images. Adv. Eng. Inf. **25**(3), 507–515 (2011)

Koch, C., Jog, G., Brilakis, I.: Automated pothole distress assessment using asphalt pavement video data. J. Comput. Civ. Eng. **27**(4), 370–378 (2013)

Li, S., Yuan, C., Liu, D., Cai, H.: Integrated processing of image and GPR data for automated pothole detection. J. Comput. Civ. Eng. **30**, 04016015 (2016)

Jog, G., Koch, C., Golparvar-Fard, M., Brilakis, I.: Pothole properties measurement through visual 2D recognition and 3D reconstruction. In: ASCE International Conference on Computing in Civil Engineering, Florida, United States, pp. 553–560 (2012)

Kyriakou, C., Christodoulou, S. E., Dimitriou, L.: Road anomaly detection and classification using smartphones and artificial neural networks. In: The Transportation Research Board 95th Annual Meeting, Washington D.C, U.S.A. (2016)

Kyriakou, C., Christodoulou, S. E., Dimitriou, L.: Detecting and classifying roadway pavement anomalies utilizing smartphones, on-board diagnostic devices and classification models. In: The Transportation Research Board 96th Annual Meeting, Washington D.C, U.S.A. (2017)

Mohamed, A., Fouad, M.M.M., Elhariri, E., El-Bendary, N., Zawbaa, M., Tahoun, M., Hassanien, A.E.: RoadMonitor: an intelligent road surface condition monitoring system. In: Filev, D., Jabłkowski, J., Kacprzyk, J., Krawczak, M., Popchev, I., Rutkowski, L., Sgurev, V., Sotirova, E., Szynkarczyk, P., Zadrozny, S. (eds.) Intelligent Systems'2014. AISC, vol. 323, pp. 377–387. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-11310-4_33

Murthy, S., Varaprasad, G.: Detection of potholes in autonomous vehicle. IET Intell. Transp. Syst. **8**(6), 543–549 (2014)

National Economic Council. The President's Council of Economic Advisers: An economic analysis of transportation infrastructure investment. The White House, Washington DC (2014)

Nejad, F., Zakeri, H.: A comparison of multi-resolution methods for detection and isolation of pavement distress. Expert Syst. Appl. **38**(3), 2857–2872 (2011)

O'Donnell, N.: Jaquar Land Rover Announces Technology Research Project to Detect, Predict and Share Data on Potholes'. http://newsroom.jaguarlandrover.com/en-in/jlr-corp/news/2015/06/jlr_pothole_alert_research_100615/. Accessed 15 July 2015

Radopoulou, S., Brilakis, I.: Detection of multiple road defects for pavement condition assessment. In: 22nd Workshop of the European Group of Intelligent Computing in Engineering, EG-ICE 2015, Eindhoven, The Netherlands (2015)

Radopoulou, S., Brilakis, I., Doycheva, K., Koch, C.: A framework for automated pavement condition monitoring. In Proceedings of Construction Research Congress, pp. 770–779. CRC, San Juan, Puerto Rico (2016)

Ryu, S.-K., Kim, T., Kim, Y.-R.: Feature-based pothole detection in two- dimensional images. Transp. Res. Rec. **2528**, 9–17 (2015)

Seraj, F., van der Zwaag, B.J., Dilo, A., Luarasi, T., Havinga, P.: RoADS: a road pavement monitoring system for anomaly detection using smart phones. In: Atzmueller, M., Chin, A., Janssen, F., Schweizer, I., Trattner, C. (eds.) Big Data Analytics in the Social and Ubiquitous Context. LNCS (LNAI), vol. 9546, pp. 128–146. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-29009-6_7

Vittorio, A., Rosolino, V., Teresa, I., Vittoria, C.M., Vincenzo, P.G., Francesco, D.M.: Automated sensing system for monitoring of road surface quality by mobile devices. Procedia – Soc. Behav. Sci. **111**, 242–251 (2014)

Yu, X., Salari, E.: Pavement pothole detection and severity measurement using laser imaging. In: IEEE International Conference on Electro/Information Technology (EIT), Minnesota, USA, pp. 1–5. IEEE (2011)

Optimization Formulations for the Design of Low Embodied Energy Structures Made from Reused Elements

Jan Brütting[1](✉), Gennaro Senatore[2], and Corentin Fivet[1]

[1] SXL, Structural Xploration Lab, Swiss Federal Institute of Technology (EPFL), Fribourg, Switzerland
jan.bruetting@epfl.ch

[2] IMAC, Applied Computing and Mechanics Laboratory, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

Abstract. The building sector is one of the major contributors to material resource consumption, greenhouse gas emission and waste production. Load-bearing systems have a particularly large environmental impact because of their material and energy intensive manufacturing process. This paper aims to address the reduction of building structures environmental impacts through *reusing* structural elements for multiple service lives. *Reuse* avoids sourcing raw materials and requires little energy for reprocessing. However, to design a new structure reusing elements available from a stock is a challenging problem of combinatorial nature. This is because the structural system layout is a result of the available elements' mechanical and geometric properties. In this paper, structural optimization formulations are proposed to design truss systems from available stock elements. Minimization of weight, cut-off waste and embodied energy are the objective functions subject to ultimate and serviceability constraints. Case studies focusing on embodied energy minimization are presented for: (1) three roof systems with predefined geometry and topology; (2) a bridge structure whose topology is optimized using the ground structure approach; (3) a geometry optimization to better match the optimal topology from 2 and available stock element lengths. In order to benchmark the energy savings through reuse, the optimal layouts obtained with the proposed methods are compared to weight-optimized solutions made of new material. For these case studies, the methods proposed in this work enable reusing stock elements to design structures embodying up to 71% less energy and hence having a significantly lower environmental impact with respect to structures made of new material.

Keywords: Structural optimization · Reuse · Stock assignment
Embodied energy

1 Introduction

1.1 Motivation

In many industrialized countries, the building sector is the most resource intensive sector [1]. Up to 50% of the total material use in Europe is associated to buildings and

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 139–163, 2018.
https://doi.org/10.1007/978-3-319-91635-4_8

infrastructures [2, 3]. Furthermore, a significant amount of energy is spent to manufacture building components as well as for the construction and the use of buildings and infrastructures. The building sector is responsible for about one third of all energy use and high greenhouse gas (GHG) emissions worldwide [4, 5]. With regard to buildings a distinction between two types of energy and carbon can be made:

- *Operational energy* refers to the energy consumption and associated *operational carbon* emissions during the use-phase of buildings (heating, cooling, etc.).
- *Embodied energy* quantifies the amount of energy used for the extraction of raw materials, the production and transport of building components as well as the building construction and end-of-life (EOL). *Embodied carbon* refers to the associated GHG emissions.

Technical standards (e.g. LEED, BREEAM, Passivhaus, Minergie) focus mainly on reducing the building operational energy. However, little attention has been given to the embodied energy, which contributes significantly to the total environmental impacts of buildings and infrastructures [6, 7]. Load-bearing systems contribute the most to the building embodied energy share [8–10] because constructing structures is a material and energy intensive process. Another issue is the construction and demolition waste (C&DW) originating from buildings. More than a third of the European waste comes from the building sector amounting to 870 million tons annually [11].

From these observations, it follows that the design and construction of buildings and infrastructures could be improved by making a more efficient use of materials. The European Commission issued an action plan to implement circular economy principles to mitigate the exploitation of natural resources, the energy consumption and the generation of superfluous waste in the building sector [12]. The circular economy paradigm suggests a closed loop flow of materials and products including recycling and reuse.

Recycling is the common approach to treat products after their use. However, recycling requires energy to process materials (e.g. melting steel) and often results in a loss of quality (down cycling, e.g. crushed concrete for road construction). An often more sustainable option is *reuse*, which avoids sourcing raw materials and requires little energy for reprocessing (e.g. reshaping the ends of a truss member to fit a new setting). Reusing structural elements for multiple service lives is identified as an opportunity to reduce the high embodied energy of traditionally designed load-bearing systems, thus lowering the overall building environmental impacts [13, 14]. A recent example where reuse was successfully implemented is the BedZED project, a large scale residential and office development in London. Up to 95% of all structural steel could be reclaimed from local demolition sites [13]. Another example is the roof of the London 2012 Olympic Stadium, which was built reusing 2500 tons of steel tubes (20% of the total structure) over ordered in an oil and gas pipeline project [5]. This adaption eventually reduced the embodied energy, construction time and overall costs. In order to assess the available elements' mechanical properties small specimen were cut out of each tube and tested [5]. Researchers at EPFL's Structural Xploration Lab [15] reused 210 skis for constructing a transportable, elastically bent grid-shell pavilion. It has been shown, that the structure made from reclaimed skis has a 85% lower cumulative energy demand than an equivalent grid-shell made of new timber.

1.2 Related Work

Structural optimization is usually carried out to improve structural performances (e.g. minimize weight and compliance) under a given set of loads and boundary conditions. Optimal truss layouts can be searched varying the geometry, topology and the member cross sections [16]. Most commonly used methods to optimize truss systems are based on the ground structure approach [17]. A ground structure consist of a grillage of many members of which only a subset will be used in the optimal design. To further improve optimality, topology and geometry optimization (i.e. varying the truss node positions) have been combined. Achtziger [18] presented methods to optimize simultaneously truss geometry and topology. Alternatively, He and Gilbert [19] employ geometry optimization to improve further optimal topology truss systems.

Generally, structural optimization is carried out using continuous design variables, e.g. member cross section areas. However, in practice only a set of standardized cross sections is available (e.g. I-beam or hollow sections) thus making the design variables discrete. Rasmussen and Stolpe [20] formulated topology and discrete cross section optimization as a mixed-binary problem including stress and deformation constraints. This problem was efficiently solved to global optimality employing branch-and-cut methods [20] and the Simultaneous ANalysis and Design (SAND) approach [21]. It was key to disaggregate equilibrium and compatibility into mixed-binary constraints that can be reformulated as linear inequalities [20].

Most optimization methods work with the assumption that structural members can be fabricated according to the optimal layout (e.g. optimal member cross sections and lengths). Usually there is no constraint on the number of available elements. Conversely, when reusing structural elements from a given stock, element properties including the cross section and element length are set a-priori and the number of elements with certain properties is finite.

An example optimization of a plane frame with fixed topology and geometry is shown in [22]. From a set of standard cross sections for each member in the structure one cross section is assigned. Each stock cross section is available only once. The objective is to minimize the weight of the structure subject to stress and deflection constraints. A genetic algorithm was employed to solve the problem. However, this formulation did not consider stock element lengths.

An arch roof built from 20 unique tree forks is presented in [23]. The stock was created through 3D-scanning available tree forks. The elements were sequentially assigned to optimally match the intended arch shape employing a meta-heuristic process.

Form-fitting strategies for constructing trusses from a set of irregular timber elements are shown in [24]. The truss geometry and topology of the static determinate system is fixed and element assignment is carried out using heuristics. The objective is to minimize waste subject to element tension and compression capacity.

1.3 Outline

Generally, in conventional structural design practice the structure topology and geometry are defined a-priori or derived via optimization methods. Then, element properties (e.g. cross section and length of a truss element) are determined. Conversely, designing a structure reusing elements from a stock entails that the element mechanical and geometric properties are given prior the structural layout is known. This is a challenging problem of combinatorial nature, which has received little attention in structural optimization.

This paper proposes new structural optimization formulations to design truss systems from an available stock of elements. The main objective is to minimize the structure embodied energy subject to ultimate and serviceability limit state constraints. Case studies are presented for three roof systems and a bridge truss. Life Cycle Assessment (LCA) is carried out to benchmark structures designed with the proposed methods against weight-optimized solutions made of new material.

2 Optimization Formulation

2.1 Assumptions for Reuse and Stock

The main underlying assumptions are that a stock of reclaimed structural elements without defects is readily available and that custom joints allow their connection. It is expected that members can be cut but not extended. A stock of structural elements is characterized by:

- Material (elasticity, strength, density, specific weight)
- Cross sections (area, area moment of inertia)
- Lengths
- Number of available elements
- Origin of the elements.

Identical elements of same material, cross section size and length are grouped.

2.2 Assignment Problem

The selection of suitable structural elements from a stock is an assignment problem of combinatorial nature. Figure 1 shows a truss layout consisting of m = 5 members. For each of the m truss member locations one element must be assigned from s = 7 stock member groups. Each group j contains n_j identical elements.

The assignment is represented through a binary matrix $\boldsymbol{T} \in \{0, 1\}^{\mathrm{m} \times \mathrm{s}}$, whose entries are:

$$t_{i,j} = \begin{cases} 1 & \text{if a stock element of group } j \text{ is assigned to the } i\text{th bar position} \\ 0 & \text{if a stock element of group } j \text{ is not assigned to position } i \end{cases}$$

The sum of all entries in each row $\boldsymbol{t}_i$ must be one to ensure the assignment of exactly one stock element at location i in the truss layout (Eq. 1). Each column of the assignment matrix corresponds to a stock element group. The sum of all entries in each column must be smaller or equal to the number n_j of available elements in the corresponding group (Eq. 2).

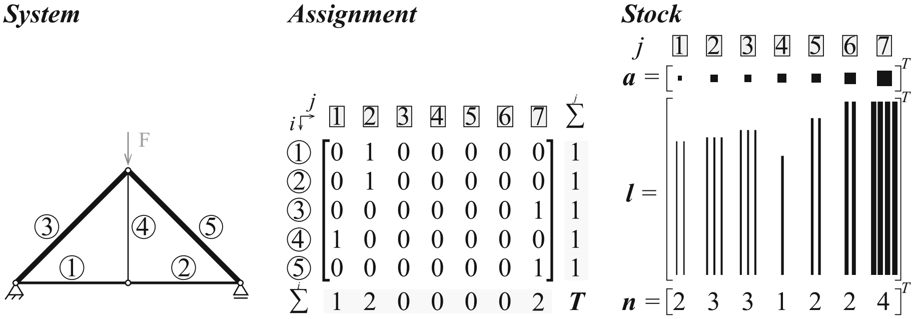


Fig. 1. Assignment of available stock elements to a structural system

$$\text{Assignment} \quad \sum_{j=1}^{s} t_{i,j} = 1 \quad \forall i = 1\ldots m \tag{1}$$

$$\text{Availability} \quad \sum_{i=1}^{m} t_{i,j} \leq n_j \quad \forall j = 1\ldots s \tag{2}$$

$$t_{i,j} \in \{0,1\} \quad \forall i,j$$

2.3 Objective and Constraints

When topology and geometry are invariant, the design variables are the components $t_{i,j}$ of the assignment matrix $\boldsymbol{T} \in \{0,\ 1\}^{\mathrm{m} \times \mathrm{s}}$. The stock is described by column vectors of available cross section areas $\boldsymbol{a} \in \mathbb{R}^{\mathrm{s}}$, Young's moduli $\boldsymbol{e} \in \mathbb{R}^{\mathrm{s}}$, area moments of inertia $\boldsymbol{I} \in \mathbb{R}^{\mathrm{s}}$, admissible stress in tension $\boldsymbol{\sigma}^{+} \in \mathbb{R}^{\mathrm{s}}$, admissible stress in compression $\boldsymbol{\sigma}^{-} \in \mathbb{R}^{\mathrm{s}}$, element lengths $\boldsymbol{l} \in \mathbb{R}^{\mathrm{s}}$, material densities $\boldsymbol{\rho} \in \mathbb{R}^{\mathrm{s}}$ and specific weights $\boldsymbol{\gamma} \in \mathbb{R}^{\mathrm{s}}$. The size of these vectors corresponds to the number of groups s collating stock elements with identical properties.

Most optimization formulations combine equilibrium and compatibility constraints. In this formulation instead, following [20], equilibrium and compatibility are treated separately. Different to [20], the formulation additionally includes self-weight, Euler buckling and the available member lengths and numbers (i.e. stock constraints).

The state variables are the member forces $\boldsymbol{p}^k \in \mathbb{R}^{\mathrm{m}}$ and nodal displacements $\boldsymbol{u}^k$ $\mathbb{R}^{\mathrm{d}}$. The vector $\boldsymbol{u}^k$ has size d which is the number of unsupported degrees of freedom. For each load case k equilibrium at nodes under external forces $\boldsymbol{f}^k \in \mathbb{R}^{\mathrm{d}}$ is computed via Eq. 4 where $\boldsymbol{B} \in \mathbb{R}^{\mathrm{d} \times \mathrm{m}}$ is the equilibrium matrix containing the element direction cosines. Self-weight is included in the external force vector adding gravity loads at the member ends through the product of matrix $\boldsymbol{D} \in \mathbb{R}^{\mathrm{d} \times \mathrm{m}}$ and the element specific weight (Eq. 4). Each column $\boldsymbol{d}_i$ of $\boldsymbol{D}$ contains half the member length l'_i at components corresponding to the vertical degrees of freedom of the element ends. The compatibility constraint (Eq. 5) relates element deformations, nodal displacements and member force via the transpose of the equilibrium matrix $\boldsymbol{B}^T$. Note that the assignment matrix is included in the compatibility equation because the element Young's moduli and cross-section areas are the inner product of $\boldsymbol{T}$ and the stock vectors $\boldsymbol{e}$ and $\boldsymbol{a}$. The operator $\circ$ indicates an element wise multiplication (Hadamard product).

Member forces are bounded by the admissible stress in tension and compression of assigned elements (Eq. 6). In addition, Euler buckling is considered (Eq. 7). Nodal displacements are bounded via (Eq. 8) to satisfy suitable serviceability limits. Only stock elements that are longer or equal to the structures' member lengths ***l'*** can be assigned (Eq. 9). Finally, the assignment and availability constraints introduced in previous section must be considered (Eqs. 1 and 2).

Objective
$$\min f\left(\boldsymbol{T},\boldsymbol{p}^k,\boldsymbol{u}^k\right) \tag{3}$$

Equilibrium
$$\boldsymbol{B}\boldsymbol{p}^k = \boldsymbol{f}^k - \boldsymbol{D}\boldsymbol{T}(\boldsymbol{a}\circ\gamma)\ \ \forall k \tag{4}$$

Compatibility
$$diag(\boldsymbol{T}(\boldsymbol{e}\circ\boldsymbol{a}))\boldsymbol{B}^T\boldsymbol{u}^k - diag(\boldsymbol{p}^k)\boldsymbol{l}' = 0\ \ \forall k \tag{5}$$

Stress capacity
$$\boldsymbol{T}(\boldsymbol{a}\circ\boldsymbol{\sigma}^-) \le \boldsymbol{p}^k \le \boldsymbol{T}(\boldsymbol{a}\circ\boldsymbol{\sigma}^+)\ \ \forall k \tag{6}$$

Buckling
$$-\frac{\pi^2\boldsymbol{T}(\boldsymbol{e}\circ\boldsymbol{I})}{\boldsymbol{l}'^2} \le \boldsymbol{p}^k\ \ \forall k \tag{7}$$

Deformation bounds
$$\boldsymbol{u}^k_{min} \le \boldsymbol{u}^k \le \boldsymbol{u}^k_{max}\ \ \forall k \tag{8}$$

Maximum length
$$\boldsymbol{T}\boldsymbol{l} \ge \boldsymbol{l}' \tag{9}$$

Optimization objectives can be functions of assignment variables, member forces or displacements (Eq. 3). Compliance optimization is formulated as the minimization of deformations under load (Eq. 10). If the volume is not constrained, this objective will result in selecting the strongest available stock elements that geometrically fit the design.

$$\min \sum_k \boldsymbol{f}^{kT}\boldsymbol{u}^k \tag{10}$$

More relevant for the case of reusing structural elements is a volume (Eq. 11) or mass (Eq. 12) minimization. Here cross section areas are minimized resulting in a better utilization of the element capacity.

$$\min V = \boldsymbol{l}'^T\boldsymbol{T}\boldsymbol{a} \tag{11}$$

$$\min M = \boldsymbol{l}'^T\boldsymbol{T}(\boldsymbol{a}\circ\boldsymbol{\rho}) \tag{12}$$

To better match the assigned stock elements with the structural layout, the total cut-off length Δl or corresponding cut-off waste ΔM can be minimized. Equation 14 weighs the element cut-off length (Eq. 13) with the corresponding cross section area, thus cutting elements with small cross sections is encouraged.

$$\min \Delta l = \sum_{i=1}^{m} \boldsymbol{t}_i \boldsymbol{l} - l'_i \tag{13}$$

$$\min \Delta M = \sum_{i=1}^{m} \boldsymbol{t}_i (\boldsymbol{l} \circ \boldsymbol{a} \circ \boldsymbol{\rho}) - l'_i \boldsymbol{t}_i (\boldsymbol{a} \circ \boldsymbol{\rho}) \tag{14}$$

2.4 Problem Nature and Reformulation

The formulation given in the previous section is a highly constrained mixed integer problem (MIP). For a given structure of m members and a given stock with a total number of $\hat{s}$ stock elements, the number of possible assignment combinations c is:

$$c = \frac{\hat{s}!}{(\hat{s} - m)!} \quad \text{where } \hat{s} = \sum_{j=1}^{s} n_j \tag{15}$$

A full enumeration of the problem is impractical even for simple structural configurations. The objective and constraint functions are linear, except the compatibility Eq. 5. This is an equality constraint and consists of a mixed binary-continuous product of assignment variables $t_{i,j}$ and nodal displacements $\boldsymbol{u}$. A reformulation of this constraint as a set of multiple linear inequalities including additional auxiliary variables is implemented as shown in [20]. This way the optimization problem is equivalently described as a mixed integer linear program (MILP) which can be solved to global optimality employing the SAND approach. SAND implies that design variables $\boldsymbol{T}$ and state variables ($\boldsymbol{p}^k$, $\boldsymbol{u}^k$) are treated simultaneously as optimization variables in the problem. The problems for the case studies presented in this paper were solved using a branch-and-bound solver [25]. The modeling of the problem was carried out with an optimization toolbox [26].

3 Life Cycle Assessment

3.1 Environmental Impacts Associated with Reusing Structural Elements

The aim of reusing structural elements is the reduction of building environmental impacts. To benchmark the impact of structures optimized with the method proposed here, a Life Cycle Assessment (LCA) is carried out. Figure 2 illustrates the main assumptions taken for the LCA of structures made from reused elements:

- Reclaimed stock elements are retrieved through selective deconstruction.
- The impacts caused through storage of elements is neglected.
- Only the elements taken from the stock are transported.
- The left over stock can be used elsewhere.
- The elements are transported to a fabrication site over a distance $d_{T,S}$ = 150 km.
- Cutting of elements happens at the fabrication site.

- The impacts caused by element ends adaption and fabrication of custom joints is neglected.
- The final structure or its parts are transported from the fabrication site to the building site over a distance $d_{T,F}$ = 50 km.
- Cut-off scrap is transported to a recycling facility over a distance $d_{T,EOL}$ = 20 km and disposed causing EOL impacts I_{EOL}.

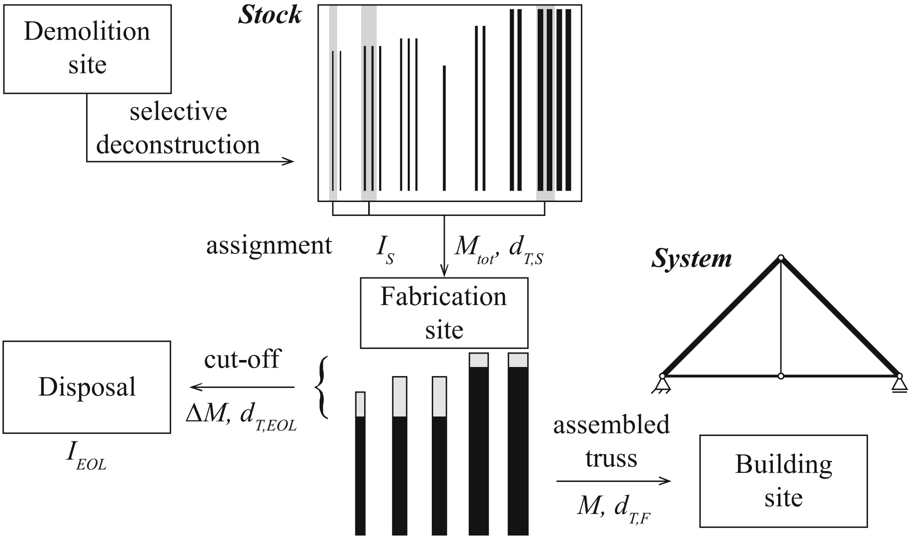


Fig. 2. Life Cycle Assessment for structures made from reused elements

The energy impacts given in [27] are considered for the calculation of impacts I_S related to sourcing structural elements for reuse. Selective deconstruction impacts I_{SD} consider a careful disassembly of a steel structure by removing connections and hoisting structural members with a mobile crane (I_C) [27]. Comparative impacts I_D caused by conventional demolition whereby all steel scrap is recycled, are reported in [27]. Because conventional demolition is assumed inevitable, for the supply of stock elements only the environmental impacts of the difference between selective and conventional demolition as shown in Table 1 are considered. The associated GHG emissions are not provided in [27] thus here converted from the combustion of diesel.

Table 1. Environmental impacts of selective deconstruction for stock element supply

Process	Energy [MJ/kg]	GHG emissions [kgCO$_{2eq}$/kg]	Source
I_{SD}	2.181	0.1878	[27]
I_D	0.359	0.0309	[27]
I_C	1.275	0.110	[27]
$I_S = I_{SD} - I_D + I_C$	3.097	0.2669	

For the impacts I_T related to transport of elements, the generic data of the Ökobaudat dataset 9.3.01 [28] is used as indicated in Table 2.

Table 2. Environmental impacts of transport

Process	Energy [MJ/(kg · km)]	GHG emissions [kgCO$_{2eq}$/(kg · km)]	Source
I_T	$0.7385 \cdot 10^{-3}$	$0.0508 \cdot 10^{-3}$	[28]

If element cutting is necessary to fit stock elements to the structural system dimensions, the off cut parts are disposed. This cut-off waste ΔM causes EOL environmental impacts in accordance Ökobaudat dataset 100.1.04 [28] which is indicated in Table 3:

Table 3. Environmental impacts of the EOL treatment of cut-off scrap

Process	Energy [MJ/kg]	GHG emissions [kgCO$_{2eq}$/kg]	Source
I_{EOL}	$12.072 \cdot 10^{-3}$	$0.8068 \cdot 10^{-3}$	[28]

From this information it is possible to quantify the total embodied energy and GHG emissions (Eq. 16) for structures made from reused elements:

$$I_{Reuse} = I_S M_{tot} + d_{T,S} I_T M_{tot} + d_{T,F} I_T M + I_{EOL} \Delta M + d_{T,EOL} I_T \Delta M \tag{16}$$

where the total mass of the selected stock elements is:

$$M_{tot} = M + \Delta M = \sum_{i=1}^{m} \boldsymbol{t}_i (\boldsymbol{l} \circ \boldsymbol{a} \circ \boldsymbol{\rho}) \tag{17}$$

Replacing M_{tot} in Eq. 16, the total embodied energy is expressed as:

$$E_{Reuse} = 3.245 \frac{MJ}{kg} M + 3.235 \frac{MJ}{kg} \Delta M \tag{18}$$

Equation (18) includes the mass M of the structure and the cut-off mass ΔM. Since mass (Eq. 12) or cut-off (Eq. 14) minimization alone could converge to different stock element assignment solutions, Eq. (18) is used to combine both objectives.

3.2 Environmental Impacts of Structures Made of New Elements

Reuse avoids the production of new structural elements. To benchmark the environmental savings through reuse against newly produced elements, common production methods involving primary and secondary (recycled) steel are considered. The impacts I_{Reuse} are compared to impacts I_{New} of weight-optimized systems. The production impacts I_P for steel MSH profiles are indicated in Table 4. This data (Ökobaudat dataset 4.1.03, [28]) is

Table 4. Environmental impacts for the production of new steel MSH profiles

Process	Energy [MJ/kg]	GHG emissions [kgCO$_{2eq}$/kg]	Source
I_P	13.175	0.921	[28]

based on Environmental Product Declarations and represents today's common manufacturing processes.

It is assumed that first-hand steel members are produced with exact lengths (no waste), having a structural mass M_{new}. The transport distance to the fabrication site is assumed to be $d_{T,N}$ = 20 km. The total environmental impacts for structures with newly produced elements are calculated as:

$$I_{New} = I_P M_{New} + (d_{T,N} + d_{T,F}) I_T M_{New} \tag{19}$$

4 Applications

4.1 Optimization of Roof Trusses with Fixed Geometry and Topology

System Description. Figure 3 shows three roof truss systems which are taken as case studies. A conventional Howe (A), Warren (B) and pitched Pratt (C) truss are considered. System topology and geometry are fixed. The span of the trusses is 12.0 m. The lateral span is assumed 6.25 m and out of plane stability is provided by other means. All member connections are pin-jointed.

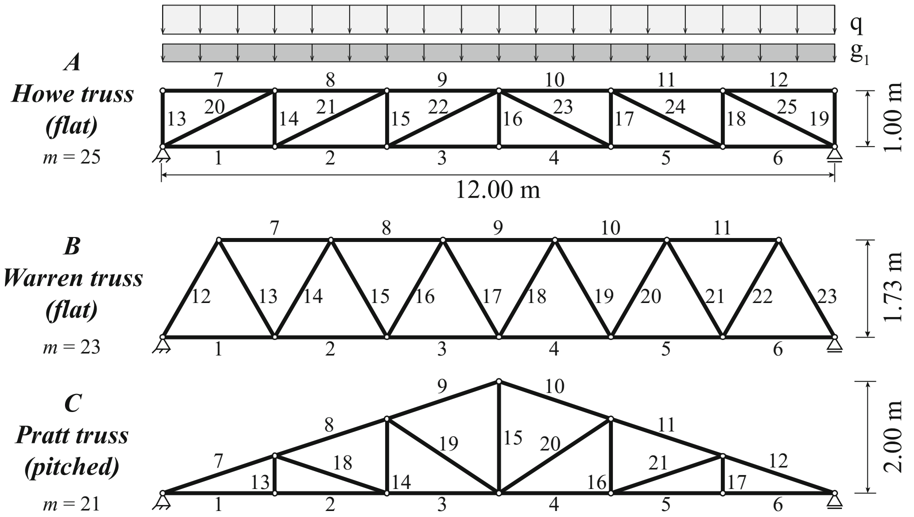


Fig. 3. Roof truss systems

Loading. The self-weight g_0 of the structure is taken into account. In addition, a superimposed dead load g_1, resulting from a roof cover and a secondary structure, is applied at the top chord nodes. Similar, a snow load q is applied over the full span. Two load combinations ($k = 2$) are considered: one for ultimate limited state (ULS) and one for serviceability limit state (SLS). For the SLS combination deflection limits are set to $l/300 = 40$ mm (Eq. 8). Table 5 summarizes load magnitudes and cases.

Table 5. Roof structures – load cases and combinations

Load case	Load magnitude	Description
g_0	From assignment	Self-weight
g_1	2.50 kN/m	Dead load (0.40 kN/m^2)
q	5.00 kN/m	Live load (snow, 0.80 kN/m^2)
Load combination	Load factors	Description
ULS	$1.35 \cdot (g_0 + g_1) + 1.50 \cdot q$	Design loads
SLS	$1.00 \cdot (g_0 + g_1) + 1.00 \cdot q$	Characteristic loads

Stock. For this study, a stock of available elements is assumed. Table 6 shows the composition of $s = 7$ element groups. The sections are square MSH profiles of variable size. All elements in the stock are grade S235 steel with a yield strength of $f_{y,d} = \sigma^+ = 235$ MPa, a Young's modulus of $E = 210000$ MPa and a material density of $\rho = 7850$ kg/m^3.

Table 6. Roof structures - stock composition

Stock group j	1	2	3	4	5	6	7
MSH	40 × 4	40 × 5	50 × 4	50 × 5	50 × 5	60 × 4	60 × 5
a_j [cm^2]	5.59	6.73	7.19	8.73	8.73	8.79	10.7
I_j [cm^4]	11.8	13.4	25.0	28.9	28.9	45.4	53.3
l_j [m]	2.50	1.50	2.00	1.50	2.20	2.20	2.50
n_j [-]	6	10	10	6	10	6	6

Comparison to Weight-Optimized Results. The three systems with reused elements are compared to equivalent designs made of new steel. In this case, the optimization is carried out assuming infinite availability of standard square MSH section types: 40 × 2.9, 40 × 3.2, 40 × 4, 40 × 5, 40 × 6.3, 40 × 7.1, 40 × 8, 50 × 3.2, 50 × 4, 50 × 5, 50 × 6.3, 50 × 7.1, 50 × 8, 50 × 10, 60 × 3.2, 60 × 4, 60 × 3.2, 60 × 4, 60 × 5, 60 × 6.3, 60 × 8, 60 × 10.

Optimization Results. The roof systems described in the previous section are optimized for three different objectives:

(1) *min M*, Mass (or volume) minimization (Eq. 12),
(2) *min* ΔM, Cut-off waste minimization (Eq. 14), and
(3) *min E*, Direct minimization of embodied energy E_{Reuse} (Eq. 18)

Figures 4, 5 and 6 show the optimal layouts of the three case study structures. In each figure the top system (a) represents the minimum embodied energy solution from reused elements. The bottom system (b) is the corresponding minimum weight solution made of new material. The member cross section size is indicated through a corresponding line thickness.

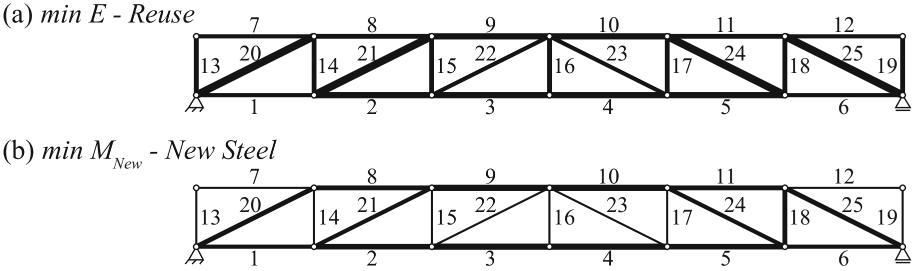


Fig. 4. System A - Optimized structures (a) made from reused elements (b) from new material

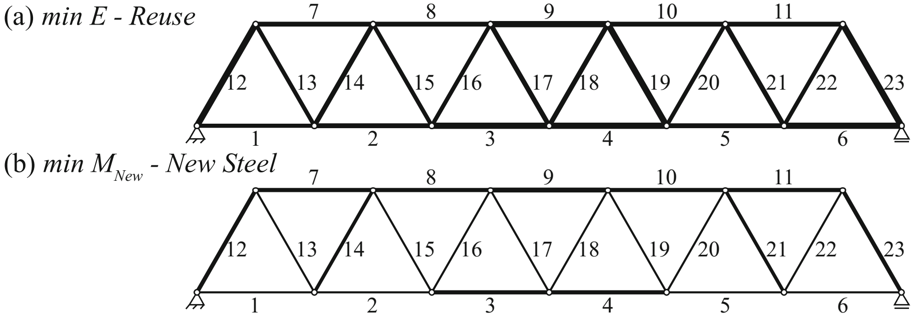


Fig. 5. System B - Optimized structures (a) made from reused elements (b) from new material

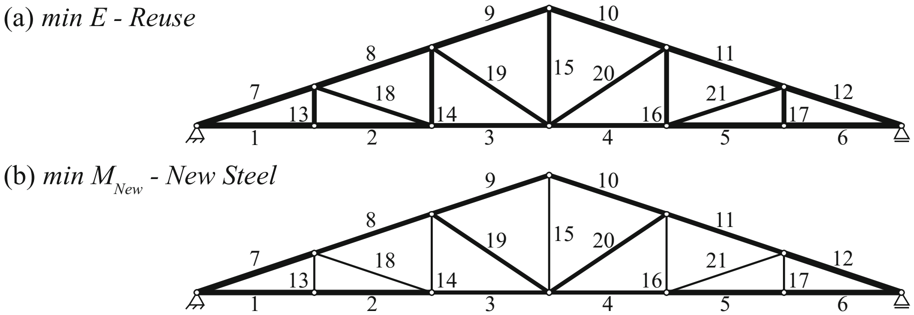


Fig. 6. System C - Optimized structures (a) made from reused elements (b) from new material

Tables 7, 8 and 9 give results for reuse systems A, B and C respectively. The right most columns give results for structures made from new material. For reuse systems A and B mass minimization is equivalent to embodied energy minimization (A: 1033 MJ, B: 922 MJ). Minimizing cut-off waste results in a slightly higher embodied energy for both systems (A: 1037 MJ, B: 1003 MJ) yet it avoids 8.8 kg and 4.9 kg of additional waste respectively. Direct energy minimization of system C results in a tradeoff between mass and waste minimization (Table 9).

Table 7. System A – Howe truss – optimization results

System A	Reuse – Objective			New steel
	min M	*min* ΔM	*min E*	*min* M_{new}
M_{tot} [kg]	318.5	319.8	318.5	–
M [kg]	**271.8**	281.9	271.8	***217.3***
ΔM [kg]	46.7	**37.9**	46.7	–
E [MJ]	1033	1037	**1033**	2874
u [mm]	28.6	28.0	28.6	31.1
util. [%]	51.1	48.0	51.1	84.6

Table 8. System B – Warren truss – optimization results.

System B	Reuse – Objective			New steel
	min M	*min* ΔM	*min E*	*min* M_{new}
M_{tot} [kg]	284.2	309.2	284.2	–
M [kg]	**261.5**	291.3	261.5	***177.9***
ΔM [kg]	22.7	**17.8**	22.7	–
E [MJ]	922	1003	**922**	2354
u [mm]	11.7	11.6	11.7	15.7
util. [%]	43.9	37.7	43.9	80.0

Table 9. System C – Pratt truss – optimization results.

System C	Reuse – Objective			New steel
	min M	*min* ΔM	*min E*	*min* M_{new}
M_{tot} [kg]	263.3	268.8	260.3	–
M [kg]	**232.8**	246.9	234.0	***200.5***
ΔM [kg]	30.5	**21.9**	26.3	–
E [MJ]	854	872	**844**	2652
u [mm]	22.7	22.6	22.7	25.7
util. [%]	64.8	55.9	61.2	86.9

The structures made from reused elements have a higher mass (up to +30%, +64%, +23% for each system respectively) than the corresponding systems made of new sections. This is mainly due to a limited availability of small cross sections in the stock (see also Figs. 10, 11 and 12) and results in a better mean element capacity utilization for the structures made of new material, compared to the reuse cases (Tables 7, 8 and 9). A comparison of the assigned cross sections for both, reuse and new material systems, are shown as bar charts in Figs. 7, 8 and 9.

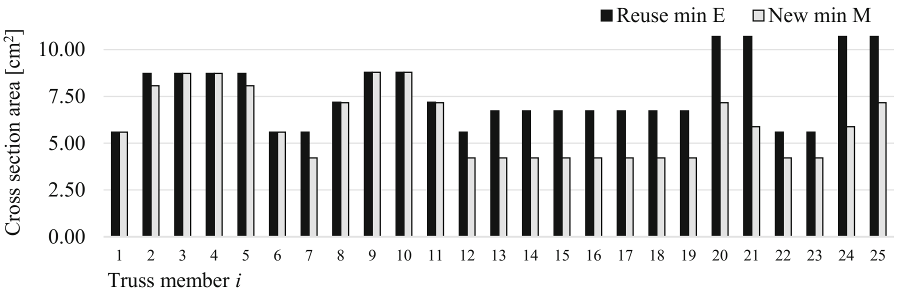


Fig. 7. System A – Howe truss – cross-section areas – reuse (*min E*) and new material system

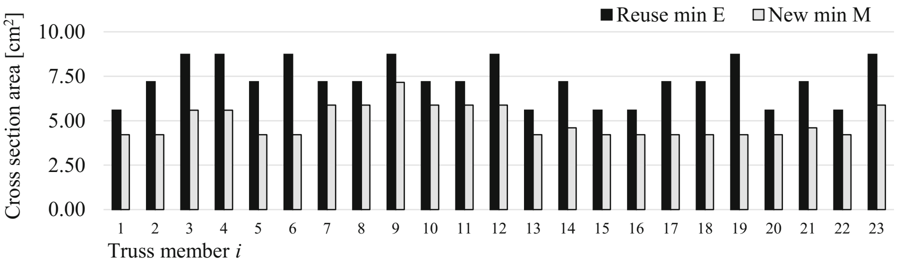


Fig. 8. System B – Warren truss – cross-section areas – reuse (*min E*) and new material system

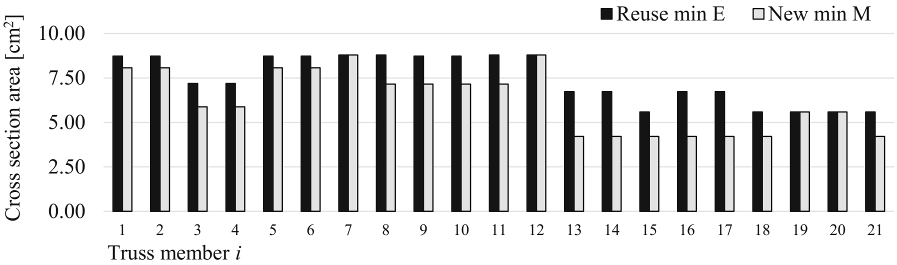


Fig. 9. System C – Pratt truss - cross-section areas – reuse (*min E*) and new material system

Figures 10, 11 and 12 show the available stock elements represented as grey bars. The elements assigned from the stock to the structures are represented by black bars superimposed on the grey bars. This way, the remaining grey part of the superimposed bars represents the length, which is cut off to fit a stock element into the structure.

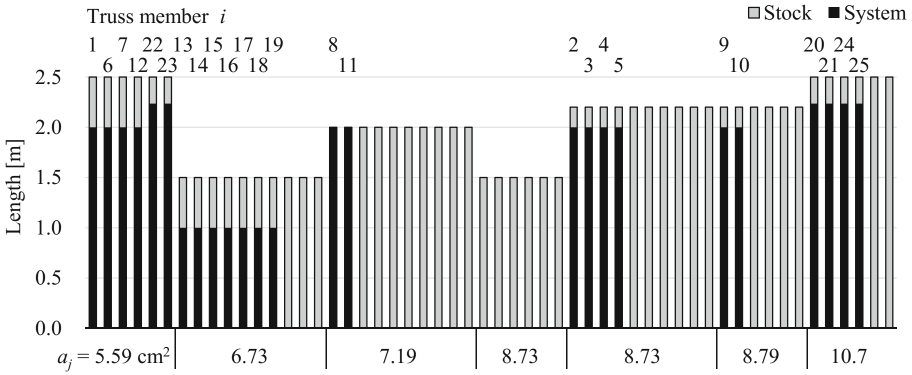


Fig. 10. System A – Howe truss – stock assignment – *min E*

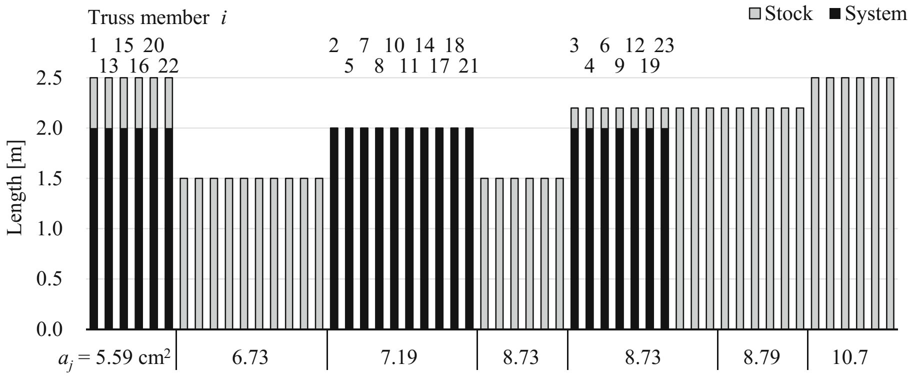


Fig. 11. System B – Warren truss – stock assignment – *min E*

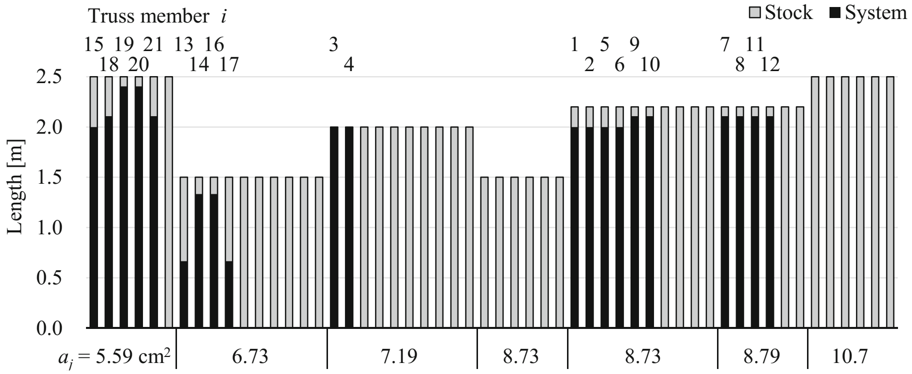


Fig. 12. System C – Pratt truss – stock assignment – *min E*

For each truss, it was possible to assign elements matching exactly the system dimensions. Elements of stock group 4 (cross-section area: 8.73 cm^2, length: 1.5 m) are not used in any structure since the elements with 6.73 cm^2 and equal length (stock group 2) have sufficient capacity or because longer element lengths were required (stock group 5).

The bar chart in Fig. 13 shows the embodied energy (left axis) and carbon (right axis) of the optimized structures for all cases.

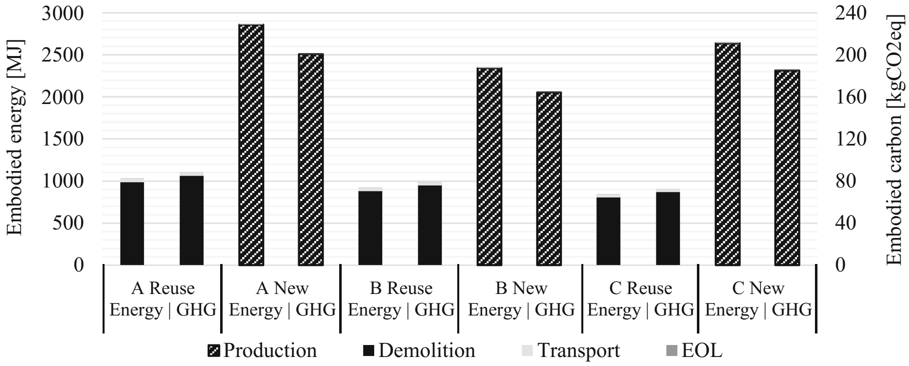


Fig. 13. Comparison of weight-optimized and reused elements structures

Transport energy and emissions are negligible compared to the production energy of new steel sections. For structures made from reused elements the selective deconstruction related energy is the biggest part of the total embodied energy. While the structures made from reused elements have a higher mass than the corresponding weight-optimized systems made of new material, they embody significantly lower energy and cause lower GHG emissions (Fig. 13). The embodied energy for reuse systems A, B and C is 36%, 39% and 32% of the corresponding weight-optimized systems.

4.2 Ground Structure Approach

Topology Optimization. In previous case studies the system topology was invariant. The ground structure approach [17] allows optimizing the topology of structures by defining a possible grillage of many bars of which only a subset will be assigned as the final structural system. To include topology optimization, the formulation is extended by adding a zero entry to each stock vector (except to the element availability $\boldsymbol{n}$), e.g.:

$$\boldsymbol{a} = \left[0, a_1, a_2, \ldots, a_j\right]^T, \boldsymbol{l} = \left[0, l_1, l_2, \ldots, l_j\right]^T, etc. \tag{20}$$

The assignment matrix $\boldsymbol{T}$ is extended by one more column accordingly. This way, it is possible to assign a *zero-element* at position i in the structure, if $t_{i,0} = 1$. This *zero-element* is $n_0 < m$ times available in the stock. When a *zero-element* is assigned, equations including stock element lengths have to be modified, such that no cut-off is added to the objective (Eq. 21) and that the length constraint (Eq. 22) vanishes.

$$\min \Delta l = \sum_{i=1}^{m} \boldsymbol{t}_i \boldsymbol{l} - \left(1 - t_{i,0}\right) l'_i \tag{21}$$

$$\left(1 - t_{i,0}\right) l'_i \leq \boldsymbol{t}_i \boldsymbol{l} \quad \forall i = 1 \ldots m \tag{22}$$

The ground structure approach might result in unstable topologies (i.e. mechanisms might arise from the assignment of a *zero-element*). It is possible to enforce the presence of members at certain locations to avoid such unstable solutions. The following constraint (Eq. 23) e.g. requires that exactly one member has to be present at either position i_A or i_B in the structure:

$$t_{i_A,0} + t_{i_B,0} = 1 \tag{23}$$

Case Study. A two span bridge truss as shown in Fig. 14 is taken here as a case study. At each side of the 4,00 m wide bridge one truss is located. Each bay of the system measures 1.80 m in length. The ground structure consist of $m = 41$ candidate bars. The number and size of the bays is chosen such that feasible assignment solutions for the given stock exist. Equation (23) is used to ensure the existence of only one diagonal in each bay, thus crossing members and unstable solutions are avoided.

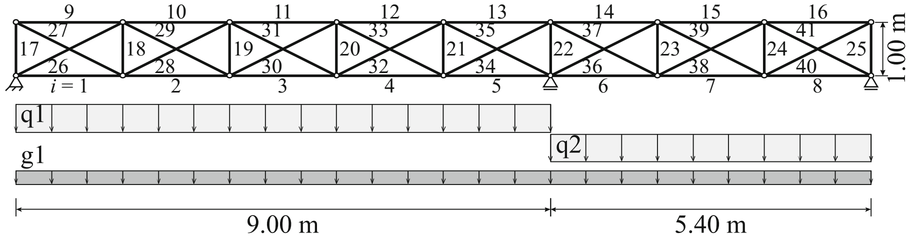


Fig. 14. Bridge truss system – ground structure

To simulate a realistic scenario, the stock of elements is assumed to originate from the selective deconstruction of the new steel Pratt truss (Sect. 4.1). Three of such Pratt trusses make up the available stock for the optimization of one bridge truss considered in this case study. The stock composition is summarized in Table 10:

Table 10. Bridge truss - stock composition

Stock grp. j	1	2	3	4	5	6	7	8	9
MSH	40 × 2.9	40 × 2.9	40 × 2.9	40 × 2.9	40 × 4	50 × 3.2	60 × 3.2	40 × 6.3	60 × 4
a_j [cm^2]	4.21	4.21	4.21	4.21	5.59	5.88	7.16	8.07	8.79
I_j [cm^4]	9.59	9.59	9.59	9.59	11.8	21.2	38.2	14.7	45.4
l_j [m]	0.67	1.33	2.00	2.11	2.40	2.00	2.11	2.00	2.11
n_j [-]	6	6	3	6	6	6	12	12	6

Loading. In addition to self-weight g_o, dead load g_1 is applied to the bottom chord (deck). Live loads q_1 and q_2 are applied to the bottom chord of the first and second span respectively (Fig. 14). Tables 11 and 12 summarize all load cases and combinations taken into consideration.

Table 11. Bridge truss – load cases

Load case	Load magnitude	Description
g_0	From assignment	Self-weight
g_1	2.00 kN/m	Dead load 1.00 kN/m^2
q_1	10.00 kN/m	Live load 5.00 kN/m^2
q_2	10.00 kN/m	Live load 5.00 kN/m^2

Table 12. Bridge truss – load combinations

Load combination	Load factors	Description
ULS_1	$1.35 \cdot (g_0 + g_1) + 1.50 \cdot q_1$	Design loads, left span
SLS_1	$1.00 \cdot (g_0 + g_1) + 1.00 \cdot q_1$	Characteristic loads, left span
ULS_2	$1.35 \cdot (g_0 + g_1) + 1.50 \cdot q_2$	Design loads, right span
SLS_2	$1.00 \cdot (g_0 + g_1) + 1.00 \cdot q_2$	Characteristic loads, right span
ULS_q	$1.35 \cdot (g_0 + g_1) + 1.50 \cdot (q_1 + q_2)$	Design loads, both spans
SLS_q	$1.00 \cdot (g_0 + g_1) + 1.00 \cdot (q_1 + q_2)$	Characteristic loads, both spans

Results. Table 13 compares the optimized structures made from reused elements against a weight-optimized structure made of new material. With regard to the structure made of new material, the same set of possible cross sections as in example 4.1 are considered. Referring to the optimization integrating reused elements, the weight-optimized truss has a mass of 189.1 kg, which is only slightly lower than that of the minimum energy solution. The structures made from reused elements are only marginally heavier (up to 16.1 kg for the minimum cut-off waste solution) than the structure made of new material. However, through the reuse of structural elements the minimum energy solution embodies only 29% of the energy used for the new material structure.

Table 13. Bridge truss – assignment optimization

System C	Reuse - Objective			New steel
	min M	*min* Δ*M*	*min E*	*min* M_{new}
M_{tot} [kg]	219.4	216.0	210.6	–
m [kg]	**189.1**	195.9	189.9	***179.8***
Δ*M* [kg]	30.3	**20.1**	20.7	–
E [MJ]	711.6	700.6	**683.2**	2378.6
u [mm]	18.3	17.1	17.8	19.0
util. [%]	64.4	56.7	59.0	64.9

Figure 15(a) shows the layout for the minimum embodied energy solution. Because the topology was optimized, the total number of bars was reduced from 41 in the ground structure to 26 members. Figure 15(b) shows the layout for the minimum weight solution obtained from new material. Both layouts use elements with larger cross section in the top chord of the left span. However, the topology differs in proximity of the central support. Figure 16 shows a bar chart comparing the used element cross sections.

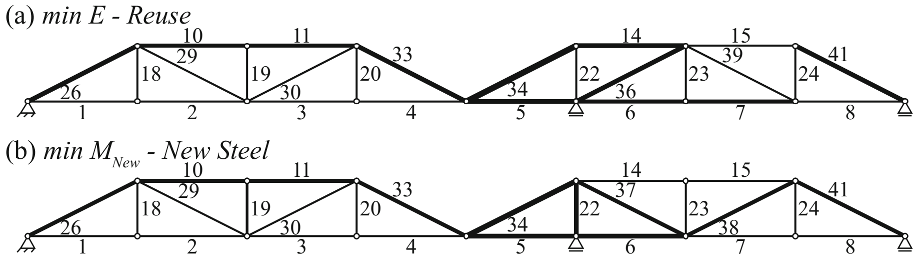


Fig. 15. Bridge truss – optimization results – final topologies

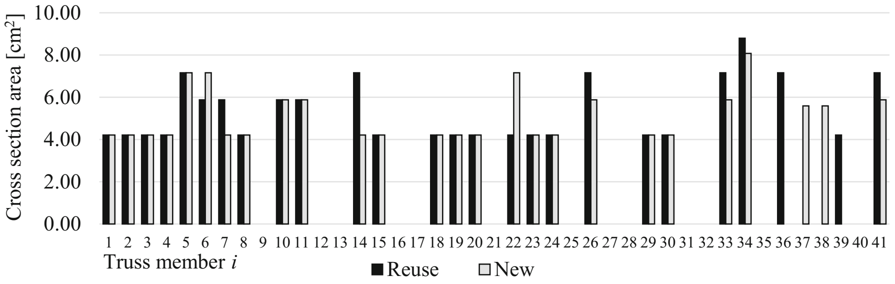


Fig. 16. Bridge truss – cross-section areas – reuse (*min E*) and new material system

The bar chart in Fig. 17 shows the stock element assignment and indicates that, for instance, some of the elements with the smallest cross-section area could not be assigned because their length was too short to fit into the structure.

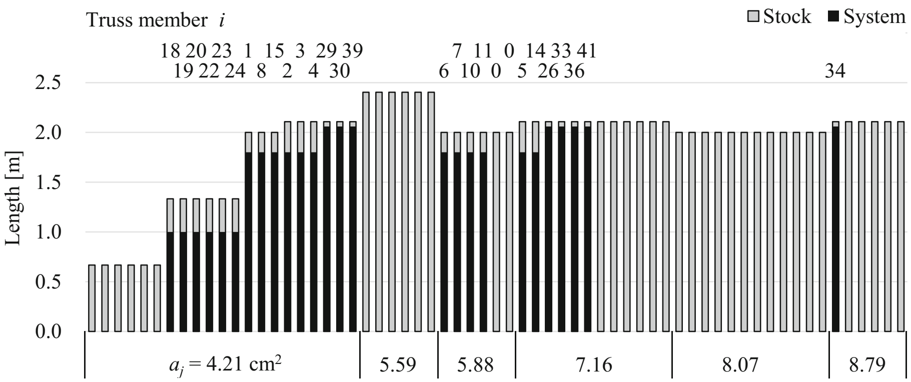


Fig. 17. Bridge truss - stock assignment – *min E*

4.3 Geometry Optimization

Formulation. In the previous case studies, the geometry of the structures was invariant. This results in most stock elements being cut to fit into the structure. However, as shown in Fig. 17 several elements assigned from the stock require only a small length adjustment. For these elements cutting could be avoided, if the structure geometry was changed, allowing the node coordinates to shift. When the node coordinates $\boldsymbol{x}$ are included in the design variables, the equilibrium matrix, self-weight and element length are functions of the nodal coordinates. All other variables remain equivalent to those introduced in Sect. 2.3. The optimization problem is given as:

Objective
$$\min f(\boldsymbol{x}, \boldsymbol{T}, \boldsymbol{p}^k, \boldsymbol{u}^k) \tag{24}$$

Equilibrium
$$\boldsymbol{B}(\boldsymbol{x})\boldsymbol{p}^k = \boldsymbol{f}^k - \boldsymbol{D}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{a} \circ \gamma) \;\; \forall k \tag{25}$$

Compatibility
$$diag(\boldsymbol{T}(\boldsymbol{e} \circ \boldsymbol{a}))\boldsymbol{B}(\boldsymbol{x})^T \boldsymbol{u}^k - diag(\boldsymbol{p}^k)\boldsymbol{l}'(\boldsymbol{x}) = 0 \;\; \forall k \tag{26}$$

Stress capacity
$$\boldsymbol{T}(\boldsymbol{a} \circ \boldsymbol{\sigma}^-) \le \boldsymbol{p}^k \le \boldsymbol{T}(\boldsymbol{a} \circ \boldsymbol{\sigma}^+) \;\; \forall k \tag{27}$$

Buckling
$$-\frac{\pi^2 \boldsymbol{T}(\boldsymbol{e} \circ \boldsymbol{I})}{\boldsymbol{l}'(x)^2} \le \boldsymbol{p}^k \;\; \forall k \tag{28}$$

Deformation bounds
$$\boldsymbol{u}^k_{min} \le \boldsymbol{u}^k \le \boldsymbol{u}^k_{max} \;\; \forall k \tag{29}$$

Maximum length
$$\left(1 - t_{i,0}\right) l(\boldsymbol{x})'_i \le \boldsymbol{t}_i \boldsymbol{l} \;\; \forall i = 1 \ldots m \tag{30}$$

Equations 24 to 30 state a non-convex, mixed-integer nonlinear problem (MINLP) for the simultaneous optimization of stock element assignment, topology and geometry of the structure. To reduce the problem complexity, a sequential approach is proposed. The outcome from the MILP formulation for stock element assignment and topology optimization is further optimized changing nodal coordinates $\boldsymbol{x}$ (i.e. geometry optimization). Geometry optimization is implemented via Sequential Quadratic Programming. Figure 18 illustrates the proposed scheme. The objective for stock element assignment with topology optimization is minimizing the embodied energy. The objective for the geometry optimization is the minimization of cut-off waste $\Delta M(\boldsymbol{x})$.

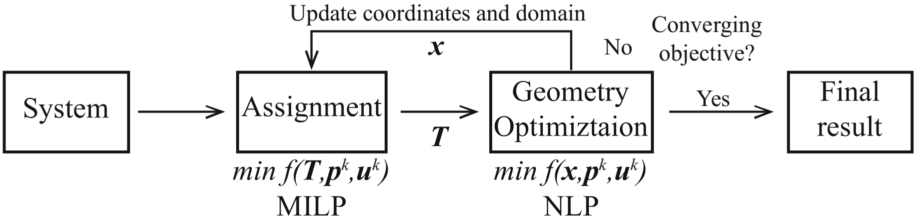

Update coordinates and domain
$\boldsymbol{x}$
No
Converging objective?
System
Assignment
$\boldsymbol{T}$
Geometry Optimiztaion
Yes
Final result
$min\ f(\boldsymbol{T},\boldsymbol{p}^k,\boldsymbol{u}^k)$
MILP
$min\ f(\boldsymbol{x},\boldsymbol{p}^k,\boldsymbol{u}^k)$
NLP

Fig. 18. Iterative assignment and geometry optimization scheme

Case Study. The same system with equivalent stock composition and load cases as shown in Sect. 4.2 is used as case study for the proposed geometry optimization. Each iteration, the coordinates of the top chord nodes are allowed to vary in a domain of ±80 cm in horizontal and vertical direction (Fig. 19). The bottom chord node positions are fixed to ensure an evenly distributed spacing and horizontal bridge deck.

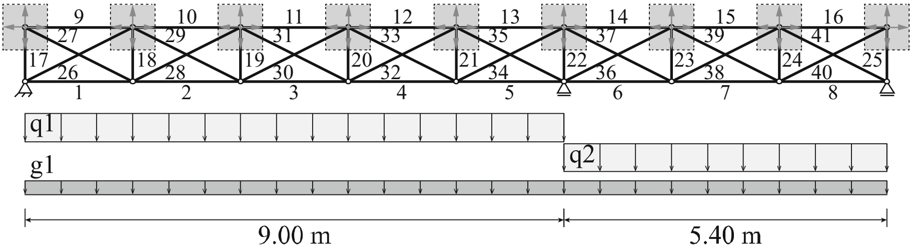


Fig. 19. System - geometry optimization

Results. Table 14 gives optimization metrics for this case study to evaluate the proposed sequential approach. The values in the column titled *1st Assignment* are those of the structure optimized in Sect. 4.2 (stock assignment and topology optimization). Geometry optimization successfully reduces cut-off waste by 7.1 kg or 34% respectively (see also Fig. 22). In addition, the embodied energy is reduced by 30 MJ (4.4%).

Table 14. Bridge truss – assignment and geometry optimization

System	1st assignment	Geometry optimization
	min E	*min* Δ*M*
M_{tot} [kg]	210.6	201.4
M [kg]	189.9	187.8
Δ*M* [kg]	20.7	**13.6**
E [MJ]	**683.2**	653.2
u [mm]	17.8	14.9
util. [%]	59.0	55.8

Figure 20 shows the optimized layout. Compared to the previous solution (Fig. 15 (a), Sect. 4.2) the top chord in the left span has changed to an arch-like shape. The right span top chord nodes are shifted upwards to accommodate the assigned member lengths. The increased height at mid-length of both spans reduces the maximum deflection compared to the previous solution without geometry optimization (Table 14).

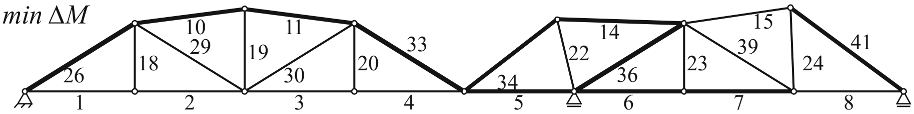


Fig. 20. Results – geometry optimization

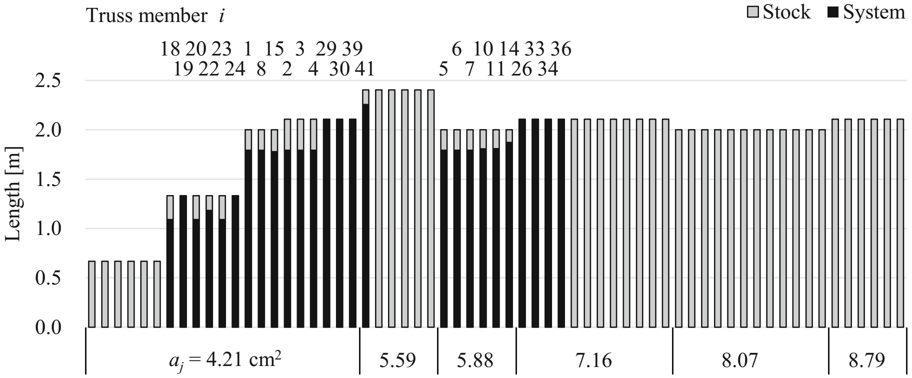


Fig. 21. Geometry optimized bridge truss - stock assignment

From Fig. 21 it can be seen that via geometry optimization nine members (19, 24, 29, 30, 39, 26, 33, 34, 36) exactly match the available stock element lengths. The variation in geometry allows the assignment of thinner cross sections, e.g. member 34 has a cross section area of 7.16 cm^2 compared to 8.79 cm^2 before (Fig. 17, Sect. 4.2).

Figure 22 shows the variation of the embodied energy, system mass and cut-off waste for successive assignment and geometry optimization. The first geometry optimization increases the weight because elements are generally lengthened. However, cut-off waste is drastically reduced by 32%. Then, the 2nd assignment carried out on the optimized geometry results in a reduction of the embodied energy and system mass. The 2nd geometry optimization slightly decreases the cut-off waste. Further assignments and geometry optimizations converge to the same solution.

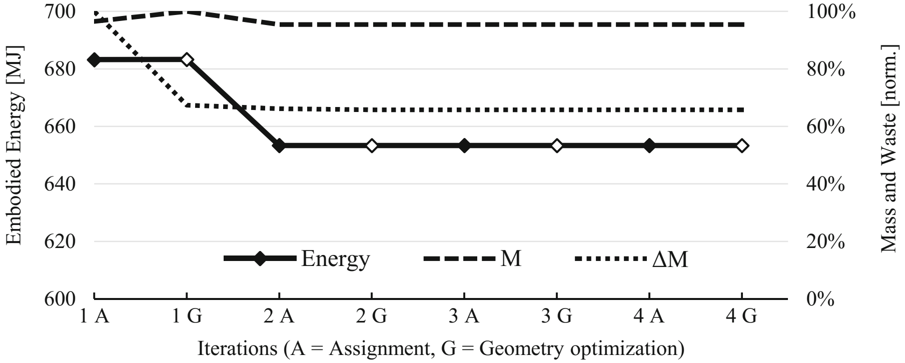


Fig. 22. Iteration history – successive assignment and geometry optimization

5 Conclusion

Reuse of structural elements offers an opportunity to reduce environmental impacts of building structures because it avoids the production of new components. This paper presented new structural optimization formulations for the design of truss structures reusing stock elements.

The selection of stock elements for a truss layout was formulated as an assignment problem, including ultimate limit state and serviceability constraints and presents a new application of discrete structural optimization problems. A sequential approach was proposed: optimal stock element assignments and topology are obtained as the solution of a mixed-integer linear program (MILP), geometry optimization is then carried out to further minimize cut-off waste. To the authors' knowledge, deterministic assignment optimization of stock elements in combination with topology and geometry optimization has not been applied yet.

A Life Cycle Assessment of the case studies taken under consideration showed that the optimized systems from reused elements embody up to 71% less energy compared to corresponding weight optimized structures made of new material. It was identified that selective deconstruction impacts contribute the most to the total impacts of structures made from reused elements.

The optimization formulation was given for truss structures. Future work could look into extending the assignment formulation to bending systems (frames, beam-column systems), which are frequently used in practice.

The proposed sequential approach for geometry optimization might result in a local optimum compared to a simultaneous assignment and geometry optimization. Simultaneous optimization of stock element assignment, topology and geometry could be investigated in next steps.

References

1. EEA – European Environment Agency: The European Environment – State and Outlook 2010. Publications Office of the European Union, Luxembourg (2010)
2. Herczeg, M., McKinnon D., Milios, L., Bakas, I., Klaassens, E., Svatikova, K., Widerberg, O.: Resource efficiency in the building sector – final report. European Commission, DG Environment, Rotterdam (2014)
3. BIO Intelligence Service: Sectoral Resource Maps. Prepared in response to an Information Hub request. European Commission, DG Environment, Paris (2013)
4. Pérez-Lombard, L., Ortiz, J., Pout, C.: A review on buildings energy consumption information. Energy Build. **40**(3), 394–398 (2008)
5. Allwood, J.M., Cullen, J.M.: Sustainable Materials: With Both Eyes Open. UIT Cambridge, Cambridge (2012)
6. Sartori, I., Hestnes, A.G.: Energy use in the life cycle of conventional and low-energy buildings: a review article. Energy Build. **39**(3), 249–257 (2007)
7. Hoxha, E., Habert, G., Lasvaux, S., Chevalier, J., Le Roy, R.: Influence of construction material uncertainties on residential building LCA reliability. J. Cleaner Prod. **144**, 33–47 (2017)
8. Kaethner, S., Burridge, J.: Embodied CO2 of structural frames. Struct. Eng. **90**(5), 33–40 (2012)
9. Webster, M.D., Meryman, H., Slivers, A., Rodriguez-Nikl, T., Lemay, L., Simonen, K., Trivedi, H., Maclise, L., Kestner, D., Bland, K., Lee, W., Lorenz, E.: Structure and Carbon – How Materials Affect the Climate. SEI Sustainability Committee; Carbon Working Group, ASCE (2012)
10. De Wolf, C.: Low carbon pathways for structural design: embodied life cycle impacts of building structures. Dissertation, MIT, Cambridge, MA, USA (2017)
11. EUROSTAT: Waste Statistics Online Database. http://ec.europa.eu/eurostat/statistics-explained/index.php/Waste_statistics. Accessed 04 Jan 2018
12. European Commission: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Closing the loop – An EU action plan for the Circular Economy, COM/2015/0614 final. Brussels (2015)
13. Addis, B.: Building with Reclaimed Components and Materials. Earthscan, London (2006)
14. Gorgolewski, M.: Designing with reused building components: some challenges. Build. Res. Inf. **36**(2), 175–188 (2008)
15. Colabella, S., D'Amico, B., Hoxha, E., Fivet, C.: Structural design with reclaimed materials: an elastic gridshell out of skis. In: Proceedings of the IASS Annual Symposium, Hamburg (2017)
16. Rozvany, G.I.N., Bendsoe, M., Kirsch, U.: Layout optimization of structures. Appl. Mech. Rev. **48**(2), 41–119 (1995)
17. Dorn, W.S., Gomory, R.E., Greenberg, H.J.: Automatic design of optimal structures. Journal de Mecanique **3**, 25–52 (1964)

18. Achtziger, W.: On simultaneous optimization of truss geometry and topology. Struct. Multidisci. Optim. **33**(4), 285–304 (2007)
19. He, L., Gilbert, M.: Rationalization of trusses generated via layout optimization. Struct. Multidisci. Optim. **52**(4), 677–694 (2015)
20. Rasmussen, M.H., Stolpe, M.: Global optimization of discrete truss topology design problems using a parallel cut-and-branch method. Comput. Struct. **86**(13), 1527–1538 (2008)
21. Haftka, R.T.: Simultaneous Analysis and Design. AIAA J. **23**(7), 1099–1103 (1985)
22. Fujitani, Y., Fujii, D.: Optimum structural design of steel plane frame under the limited stocks of members. In: Proceedings of the RILEM/CIB/ISO International Symposium, Integrated Life-Cycle Design of Materials and Structures (2000)
23. Mollica, Z., Self, M.: Tree Fork Truss. In: Adriaenssens, S., Gramazio, F., Kohler, M., Menges, A., Pauly, M., (eds.): Advances in Architectural Geometry 2016, pp. 138–153. vdf Hochschulverlag, Zürich (2016)
24. Bukauskas, A., Shepherd, P., Walker, P., Sharma, B., Bregula, J.: Form-Fitting strategies for diversity-tolerant design. In: Proceedings of the IASS Annual Symposium, Hamburg (2017)
25. Gurobi Optimization Inc.: Gurobi Optimizer Reference Manual. www.gurobi.com. Accessed 04 Jan 2018
26. Löfberg, J.: YALMIP: a toolbox for modeling and optimization in MATLAB. In: Proceedings of the CACSD Conference, Taipei, pp. 284–289 (2004)
27. Athena Institute: Demolition energy analysis of office building structural systems. The Athena Sustainable Materials Institute, Ottawa (1997)
28. Ökobaudat – Datenbasis für die Ökobilanzierung von Bauwerken. Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, Berlin (2017)

Collaborative Immersive Planning and Training Scenarios in VR

Christian Eller(✉), Timo Bittner(✉), Marcus Dombois(✉), and Uwe Rüppel(✉)

Technische Universität Darmstadt, Franziska-Braun-Str. 7, 64287 Darmstadt, Germany
{eller,bittner,dombois,rueppel}@iib.tu-darmstadt.de

Abstract. Virtual reality is a promising technology to create fully immersive exercises in various fields such as civil safety engineering, security, teaching or facility management, designed as serious games for engineering purposes. Based on previous research and cooperation with industry partners and public institutions a concept for a collaborative virtual environment for multiple organizations with different hardware systems was devised. The approach focused on the integration of 1st person virtual reality together with the corresponding interactions as well as a roles and rights management for all participants involved. Additionally, the setup allows for traditional desktop solutions to be included and enables users to build a dynamic and ever-changing training scenario based on an interactive planning phase. The concept was implemented and tested within a specific civil safety scenario using modern head-mounted displays as well as controllers and object tracking capabilities. This paper outlines the necessary steps in order to achieve such a collaborative virtual environment including various responsibilities for different roles and hierarchy levels and examines its potential. Together with the use of different hardware systems the described approach results in a flexible concept for planning and training applications in order to create dynamic, individual scenarios that facilitate the effectiveness of learning.

Keywords: Virtual reality · Civil safety engineering
Collaborative virtual environment · Scenario planning and training
Serious games

1 Introduction

In the field of security and safety practical trainings have always been used to exercise organizational processes and to strengthen the problem solving skills of the individual, the whole team and across organizations. A recurring problem with the trainings, however, is that especially large trainings with participants from different organizations cost a lot of organizational and financial effort. Also, due to structural or economic issues, for example, the trainings often cannot be carried out everywhere where it would make sense. An example of this can be the mass casualty incident exercise carried out by the fire brigade and a number of rescue services in Southern Brandenburg (Germany) in 2017 [1]. The exercise involved about 400 people from different organizations and took

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 164–185, 2018.
https://doi.org/10.1007/978-3-319-91635-4_9

almost a year to be organized. In order to minimize the number of overlaps with daily operations, the exercise held on hospital grounds was carried out on a Saturday.

But not only the field of security and safety, specifically emergency response and crisis management, could benefit greatly from taking practice to the next level. Various concepts already exist to improve the results by using modern technology and applying new methods to facilitate learning success. As such, a lot of aspects regarding planning and training in various fields like education, manufacturing and production, facility management and many more can be improved upon by utilizing a more hands-on approach that is as close to reality as possible. This is where virtual reality (VR) comes into play – literally. Pushed by recent advancements that mostly stem from the gaming industry current VR applications that make use of head-mounted displays (HMDs) and advanced tracking methods hold great potential in regards to transferring real movements into a virtual environment that offers intuitive visualization. Properly utilized this can lead to better results as the processes involved engrain themselves deeper in the trainee's mind and muscle memory. And while areas such as sports can mostly replicate the appropriate situations during training, there are a number of fields that need to find a solution elsewhere, be it that an adaptation of realistic scenarios would require too much space, be inadequate, too costly or straight up impossible to carry out. New developments allow multiple users to dive into a virtual environment at the same time all the while using different technologies and even tracking and incorporating real world objects into a given scenario, even combining 1st person VR and desktop applications. Following this trend, the enhanced immersion from VR coupled with other approaches can lead to positive results as has been established within the scope of so called serious games.

This paper aims to further examine the potential of virtual reality applications in the given context. The use of replicating the real world within a virtual environment that offers various possibilities to practice whatever situation is required or desired has been a subject of research for several years now and modern technologies coupled with a broader availability of the necessary hardware open up plenty of opportunities for new and improved planning and training methods. Immersive and interactive scenarios can be used to increase the effectiveness of learning: trainees not only observe or listen but rather play an active part by influencing the digital environment around them. Here, VR provides the framework for a highly customizable application that is tailored towards the specific needs at hand and may go beyond traditional training methods in order to study and absorb motion sequences, build muscle memory and impart knowledge on the participants. Employing collaboration techniques and multi-user environments in this context provide a wide range of possibilities, i.e. through the use of student-teacher interaction or cooperative efforts by working on a task together. The approach established here is based on feedback of emergency response teams and institutions as well as a literature review of the related work. The collaboration aspect is implemented through different roles and rights, for example by allowing only certain participants to perform a given task or manipulate specific objects. This design enables a number of interactions and behaviors that can be utilized to carry out the designated scenario by switching between active and passive roles and views. The focus further lies on 1st person VR and interactions enhanced through real world objects as well as in-game scenario building for dynamic exercises and adaptive learning.

The approach was implemented featuring a fire drill that includes both a tactical planning phase and a training phase. Roles like fire departments chiefs and squad leaders can guide through the first phase using several tools that are only at their disposal. Decisions and changes made during briefing will influence the events within the second phase, where the assignments for those that were observers will change as they now have to translate the results into actions, i.e. by putting out fires and following other orders. This exemplary application will be implemented via available modern VR technology such as HMDs but also support other means of input like keyboard and mouse together with a regular monitor. This allows for dynamic custom-fit solutions that improve existing practices and learning concepts while enabling detachment from the aforementioned spatial, monetary and organizational constraints. The collaborative virtual environment holds great potential when the technical possibilities mentioned above are applied to the full extent to achieve an immersive and holistic approach which can increase the effectiveness of learning.

The paper is organized as follows: Sect. 2 gives a short introduction to virtual reality and an overview about the related work; Sect. 3 contains the description of the requirements; Sect. 4 exemplifies the concept and design for the collaborative virtual environment including the roles and rights management; within Sect. 5 the implementation and the scenario are explained. And finally, Sect. 6 draws the final conclusions and gives some input regarding outlook.

2 Background and Related Work

Virtual reality as a field of research is relatively new and highly correlates with the development of the necessary computer hardware. Up to now there is no uniform definition for it. On the one hand it is a technology that enables the creation of immersive and interactive environments using special hardware and on the other hand it refers to a methodology for an inclusion of users to a seeming reality [2]. Compared to conventional computer graphics VR allows a multimodal presentation including visual, acoustical and haptic components. Presentations, interactions and simulations in VR are time critical and have to be done in real time. Whereas in conventional computer graphics interactions are usually done in a two dimensional manner (e.g. keyboard and mouse) and the perspective is user-independent, VR uses 3D-interactions including movement tracking (e.g. head and hand tracking) and allows for a user-dependent perspective [2].

As described by Guillen-Nieto and Aleson-Carbonell [3] the value of VR for education and training purposes within serious games has been shown multiple times in research and in practice. Next to the effect of an increased personal motivation and satisfaction, different styles of learning are accommodated for and decision-making as well as problem-solving is being supported. Serious games and VR also have a positive impact on the effectiveness of learning outcomes as shown by the authors. Especially virtual reality-based instructions and task planning [4] resulted in a higher learning outcome. The use of VR has furthermore been proven to be effective for collaborative learning for example in e-learning environments [5] and other fields [6].

Based on the mentioned learning results, it is not surprising that VR is currently being used in many areas for learning, training and planning purposes: In [7] for example the authors implemented a virtual prototyping system to allow the collaborative planning and design of vehicles. They argue that "virtual prototyping is a logical extension of computer-aided design…" and helps to overcome the problem of having people from multiple locations working on the same project. Another example is shown in [8] where the authors among others designed the "Sky Classroom", a VR-tool that aims to teach global engineering collaboration skills. The avatar-based platform allows students to virtually enter the building construction site and review the BIM model.

Potentially the biggest field where serious games and VR are used is the area of security and safety including medical and military trainings as well as civil safety and emergency management. An example is the modular virtual reality patient simulation system for U.S. Navy medical providers [9] that offers both, a three-dimensional VR and a two-dimensional desktop interface for teaching and training of cognitive assessment and treatment skills. Pucher et al. [10] argue that mass casualty incidents have become more common and therefore low cost, immersive VR platforms for training and assessment could represent "[…] a powerful adjunct to existing training methods […]". They designed, implemented and evaluated a virtual clinic where multiple doctors could simultaneously manage the situations within the limits of the hospital's resources. Sharma et al. [11] propose a collaborative virtual environment for preparedness and response training in subway evacuations. Based on *Unity3D* and *Photon* they designed a multi-user avatar-based training environment in the cloud. Through the combination of computer simulated and user controlled (using provided controls or a joystick and a VR-headset) agents, they try to analyze crowd behavior and people trust in evacuation scenarios. Passos et al. [12] provide and evaluate a collaborative virtual environment and compare it against commercial solutions and other research projects. They use *Autodesk 3ds* in combination with *Unity3D* to simulate big events and train security agents. Their research focusses mostly on the collaboration model including communication, coordination and cooperation and the scenario design tools. Like before they use avatars controlled by users with keyboard and mouse to play the serious game. Some tools for data analysis were included for a better scenario evaluation.

Next to the mentioned publications and finished research projects there are also examples for currently ongoing projects like the Auggmed project [13] or commercial solutions like the XVR platform [14] or LUDUS [15]. They all focus on scenario building and training for collaborative multi-user scenarios in security and safety. In addition to that they all include features for the evaluation and analysis of the training scenario.

The use of serious games and VR for planning and training scenarios as described in the publications and systems mentioned above is already proven to be an effective tool for learning. Especially the design of collaborative virtual environments and multi-user scenarios was shown to be of high importance for most learning areas. In addition, the implementation of these systems in a cloud- or server-based fashion is an important prerequisite as many teams do not share the same spatial location. To allow scenarios that feature for example big crowds it is appropriate to combine computer simulated avatars with avatars controlled by the trainees. To be able to evaluate the learning experience it is furthermore worthwhile to add tools for analysis in regards to the developed system. In most papers much effort goes into the scenario building and avatar design. However, so far all approaches focus on agent based trainings where the trainee uses a joystick or a keyboard plus a mouse to interact (sometimes in combination with a VR-headset for head-tracking). This results in a non-immersive training scenario and mostly 3rd person visualizations and interactions. Moreover, the number of interactions might be limited due to predefined, static scenarios and real world objects are not included in the virtual environment. What is needed is a generic virtual training environment that allows users from different organizations or acting within different roles to directly interact with their surroundings and each other in a 1st person manner and include real world objects into a dynamic scenario in order to facilitate the immersion and enhance the learning experience and outcomes.

3 Requirements for the Collaborative Virtual Environment

While there are many successful applications considering the training of individuals or teams for tasks and missions in the field, the planning aspect is often viewed separately or disregarded entirely. Thus, the interaction between user and environment as well as the collaborative aspects mostly take place in a set scenario with changes stemming mostly from user interactions that are different between repetitions while the surroundings and tasks remain the same. In order to overcome those inherently static approaches the potential of incorporating the results from planning was identified as a way to introduce variety and improve learning effects. The added value that arises from the inclusion of planning aspects for the training of emergency forces has been known for some time and is taken up in these areas by means of simulation games on stationary training boards [16]. Other examples also show that situation analysis and planning in crisis management has already established itself as a great tool and is practiced by various means, such as the use of interactive table tops [17]. The following section identifies the requirements with regard to the use of VR for planning and training in accordance with the method of requirements engineering. This method is comprised of identifying, analyzing, specifying and validating the characteristics and surrounding conditions of the software system [18] and allows researchers to work closely with experts, programmers and operators alike. Aspects of planning are incorporated with the aim to develop an immersive, interactive and holistic experience for multi-user VR based on the correlating procedures. It is of significant importance to clearly define all the requirements that have to be considered when designing a concept for a collaborative virtual environment that carries out dynamic scenarios.

The primary component that was identified in regards to a collaborative virtual environment is by definition that of multiple users in the same scenario. However, in the given context of utilizing planning to make the scenarios dynamic in nature it is no longer sufficient to simply rely on the for the most part limited scope of conventional networking solutions. While there is a plentitude of possibilities for basic operations such as establishing connection between clients or sending messages and other information, the implementation of the planning aspects and the various interactions that come with it needs to be much more refined. This includes but is not limited to the management of roles and rights for the manipulation of users and objects. For example, if multiple organizations are planning a scenario it has to be ensured that clients can only interact with things they are allowed to while maintaining a unified view for all parties involved. Due to that, the environment's implementation has to allow for a multitude of clients with differing access rights that can be granted and revoked at runtime. In order to account for those changes, all objects relevant to planning have to be integrated in the networking infrastructure one way or another. The roles and rights management component is required to handle both a fixed hierarchical command structure that affects interactions as well as a flexible system that features things like ownership requests and has an according and intuitive representation in the environment.

While handling different levels of access and interaction in the various scripts that are responsible for objects and actors the visualization is also an important requirement in order to ensure intuitive controls and understanding. If this is not provided, such a complex system may quickly become an intrusion or a distraction when inside the virtual environment and hinder planning and training alike. The requirement is therefore set to have colors that accurately represent whichever organization the object belongs to and gives the user an idea of the possible interactions. Since this approach is limited in a couple of ways additional help should be provided, for example through the use of meaningful and easy to understand names and variables that will show up on the user interface (UI) or the information board in the meeting room during the planning phase. This extends to the avatars of the users within the virtual space. Not only should they reflect either a certain person or role, but they should also be accurate enough so that actors will know immediately whether or not they are talking to or interacting with another party. An additional requirement is thus formed to outline a character representation concept that defines the visual aspect of user avatars. Since modern HMDs like the *HTC Vive* or the *Oculus Rift* rely on the tracking of the head as well as two controllers (one for each hand) it became immediately obvious that those will also be the core components of character representation. Since full body avatars are not only highly complicated to animate in a way that does not impair user experience due to inaccurate or flat out wrong movements in combinations with factors like clipping but also have to be networked correctly within the collaborative scenario, the requirements for avatars where reduced to representing the position of head and hands. Clients receive both the position and the rotation which still allows for an adequate perception of other users and their movements and interactions. For example, one of the most basic but also most important requirements was the look direction that enables an immediate understanding of what other actors are currently doing or talking about. Coupled with tracking and networking the controllers that should not be limited to the visual representation of hands but also allow for various tools such as pointers and markers the rather abstract and simplistic approach to avatars is deemed sufficient. An additional benefit is that this requirement also prevents users from obstructing each other's views when huddled together in the meeting room. Concluding these specifications, the environment should seamlessly provide integration for all possible interactions as well as character representations (avatars) and offer an intuitive surrounding that holds all necessary features and objects of the planning and training scenario. While a certain level of visual quality is definitely not only desired but also needed for a sufficiently immersive environment and intuitive handling of objects photorealism could never be an actual requirement. Not only are current HMDs limited in quality, e.g. due to resolution, but implementing such high quality visuals in the given context is neither time- nor cost-efficient. Furthermore, once computer graphics reach a level like that, the so called uncanny valley may lead to unpleasant side effects [19]. Thus the overall requirement for environment visualization can be described as a trade-off between quality and performance, basically stating that as long as the user is not occupied with graphical or other shortcomings of the virtual environment the specifications are adequate.

Following a similar line of not interfering with immersion and facilitating intuitive comprehension are the requirements for the focus on 1st person VR. While allowing spectators and even certain actors to use regular desktop applications the main goal is the integration of intuitive 1st person interactions through HMDs and controllers as well as real world objects that enhance the learning effects of planning and training. This includes, but is not limited to, different views and camera perspectives that improve user experience by enabling them to follow the action in a both comfortable and practical manner. Controller haptics and interfaces for different hardware and application types are of equal importance when it comes to providing a suitable VR environment. This should lead to an implementation that features easy to grasp manipulation of objects, for example in the sense of what can be grabbed and how, as well as more abstract interactions like granting or requesting permission for certain actions and working together with other users. Lastly, this requirement also entails the use of real world objects that are tracked and networked similar to the HMDs and controllers, currently available for the *HTC Vive* [20]. As such the specifications for 1st person integration should be able to handle different HMDs, incorporate tracking and interaction with both virtual and real world objects that is intuitive and improves the multi-user scenarios by processing all necessary information within the network. Since the scope of this requirement remains fairly wide it will be explained further in Sect. 4.

As not every interaction can be carried out through the means of what is inherently provided by the VR systems, custom tools are another critical requirement in order to make planning and training work in a collaborative virtual environment. This includes a lot of features that can be found in most VR applications like teleportation for when users run out of tracked area or want to participate while remaining seated at their desk. The requirements for movement in the virtual space also extend to aspects like limiting users to certain areas and providing tools like navigating to a given target. Together with the avatar representation described above this ascertains that users can follow others while moving fast or even teleporting. During the planning phase, the UI has to implement several context-sensitive features that pertain to manipulating the relevant objects but also administrative operations like adjusting scenario settings, managing roles and rights and finally initiating the training phase that is based on the dynamic features of planning. One example of this is the placement of so called pawns. Similar to chess pieces, those pawns represent certain aspects of the training scenario that can be moved around to place spawn points (determining where an organization or user begins the training scenario), location markers (used for pathfinding or highlighting tasks) or influence the environment by blocking pathways or diversifying tasks depending on the desired scenario. Specifically, the collaborative virtual environment should make it possible to practice the same situation in a multitude of different ways. Again, a core aspect of the requirements is to keep them as flexible as possible in order to be able to design varying components that facilitate different learning goals. Upon initiating the secondary phase, pawn placement together with all the other features and tools is transferred into the training environment, thus creating a dynamic scenario based on the results of the planning phase. It has to be considered that a lot of those tools not only have to be made available for individual users but also need to be networked properly and are potentially influenced by other actors. For example, the leader of an organization can grant and revoke certain rights for other members during the planning phase, e.g. letting them interact with specific objects, or direct and observe them during the training phase while changing environment parameters.

All aspects discussed in this section aim to support an implementation that remains as generic as possible to allow for various types of scenarios that are dynamic in nature. In order to accomplish that, the definitions above ensure that the proposed system remains independent of location and technology. This is achieved by using online networking that supports a multitude of users and organizations, different platforms as well as application types and provides tools for dynamic scenarios in various contexts. Since the possibilities of integrating such tools in a 1st person environment are basically limitless, the corresponding requirements merely outline necessary features and considerations without narrowing down the conceptual implementation which will be exemplified in the following sections.

4 Concept and Design of the Collaborative Virtual Environment

Based on the requirements outlined in the previous section, a concept for a collaborative and interactive planning and training scenario in the scope of serious games was developed. Important for this concept is a modular structure to transfer possibilities and opportunities easily in different scenarios and fields of applications like teaching, facility management or rescue training. Although the implementation should be transferable and enable individual solutions, each scenario in a distinct field also offers various outcomes through use of planning and training. In the planning room, there is the opportunity to discuss the action before training with all users and influence the scenario individually, for example by introducing new goals, starting positions and obstacles to the scene. That way the focus can be changed in reoccurring exercises or to add something new and unknown. The concept shows the usage of VR for a fully immersive scenario while retaining the option of using a regular desktop application for an easy overview without the need for an HMD. This section will mainly focus on the aspect of building a generic but dynamic multi-user environment. The first part outlines how the project-specific components like the *Unity3D* engine and different software development kits (SDKs) will work together to create the baseline for the collaborative planning and training. Subsection 4.2 explains the use of different objects by various users and roles.

4.1 Modular Components in Unity3D Using Photon and VRTK

In order to create a scenario which is ready for both VR and non-VR systems and usable independent of location by means of a networking solution a game engine called *Unity3D* [21] was used. *Unity3D* is an extensible and established engine via which flexible scenarios can be created based on a modular programming paradigm. An active developer group is steadily expanding the functionality and possibilities of the engine by developing SDKs or packages with objects, textures, scripts or entire modules which can be easily integrated into a project. Therefore, *Unity3D* is a game engine which is highly customizable and due to the structure of tailored to fit packages it allows for development of transferable scenarios by means of so called game objects. *Unity3D* also combines graphics and functionality through an easy import from a large number of file formats that can be implemented through the aforementioned game objects, leading to a hierarchical and accessible project structure and an extensible implementation. Since *Unity3D* is limited in terms of modeling, it is recommended to rely on *Autodesk 3ds* or similar software to develop the objects visuals or the environment as a whole and import it into *Unity3D* [12]. The resulting 3D models can be of various file formats such as *FBX* which can be exported from numerous programs, even utilizing entire building geometries available through building information modeling (BIM) software overhauled in *Autodesk 3ds*. The so called mesh of an object visualizes the underlying information in combination with textures to create the environment for a scenario. Within the *Unity3D* project the objects obtain functionality through scripts. Due to the hierarchical and modular structure game objects can be transferred to another project or scenario while retaining all features when developing a collaborative virtual environment for dynamic planning and training scenarios.

One of the required extensions is the software *Photon* which is a tool to implement a networking solution and automatically manage online lobbies via a master-client relationship, sending data to all participants [22]. *Photon* provides a free cloud-based option which is able to manage a multitude of users in one virtual room and distribute the data of the scenario. Every user has a local copy of the scenario which is linked to the network lobby. The lobby connects the users and the virtual room so that everybody can send or receive data through the lobby. Sending data is limited to 400 messages per second, meaning that it is important to control the exchanges when the number of users increases and reduce the data sent if necessary. For the intents and purposes described here however, the limits are high enough to be neglected [23]. The incoming data is processed through the use of various scripts and can be send to specific users, groups or everyone in room, directly influencing user interaction via a process that is called interest management. As such, the *Photon* engine is a satisfactory solution to network the collaborative scenario (see also [11]).

Another useful package is the virtual reality toolkit (VRTK) [24] which includes extensive functionality via customizable scripts to simplify a VR solution in *Unity3D*. With *VRTK* a VR user can be created that handles tracking of the HMD and the controller. Examples include scripts to enable interactions with a controller or the HMD in order to touch, grab or otherwise manipulate objects as well as interacting with an UI to control buttons, input fields or check boxes. *VRTK* establishes connection with *Steam*

VR, an SDK used for most HMDs to manage the tracking of the hardware, representing and transferring tracked objects in the virtual world and transmitting information such as user inputs, e.g. a click on the trigger. Therefore, *VRTK* can be used to handle the *HTC Vive* and the *Oculus Rift*, but also to instantiates a simulator if no appropriate HMD is found. The simulator can be treated similar to any 1st person VR user, although interaction is achieved via keyboard and mouse only which is less intuitive and does not build muscle memory like HMD and controllers. Figure 1 shows the structure of the usage *Unity3D*, *VRTK* and *Steam VR* with a *HTC Vive* or *Oculus Rift* HMD. As described in the requirements it is important to enable different points of view to be able to follow the action in the scenario. This is done with a spectator view which can be controlled with common input devices like keyboard or joystick allows for the use of a regular desktop application. An additional view together with the required tools is implemented for instructors who lead the training, give feedback or hints for specific situations and problems to participants and interact as a neutral user. This allows observer and instructors alike to utilize the various views in order to analyze and judge characteristic values that are not measurable, for example the ability to work in a team.

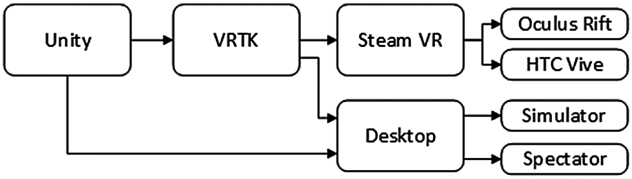


Fig. 1. Implementing different view modules and hardware systems

Due to the chosen design it does not matter if the participant is using a *HTC Vive* or *Oculus Rift* for 1st person VR, moves via "real movement" in a tracked area of 5 × 5 m, teleports to be independent of real-world boundaries, e.g. while sitting down at a desk, or uses the trackpad or joystick to walk and control without an HMD. Another alternative would be a so called locomotion platform, e.g. *Virtualizer* [25], to move without boundaries through the entire virtual room. With the help of *Photon* and *VRTK* an avatar is sent to the users so they can perceive position and movement of other participants no matter the view or hardware. The described procedure also employs the modular approach mentioned above and can thus be transferred to other areas and applications. A combination of *VRTK* and *Photon* can be used to create a networking script that handles the collaboration, interaction and planning in one room while the framework and the environment is handled in *Unity3D*. It is also possible to spectate other users depending on the chosen hardware and perspective. The resulting game objects in *Unity3D* may be transferred to different scenarios and can be turned on and off as required.

This further exemplifies the modular project structure explained above and enables users to individualize the scenario before training by adding, manipulating or removing objects. It is also possible to deactivate whole parts in a scene and enable them again when they are needed, distinguishing between purely visual features or full-fledged functionality. If there are no direct dependencies between the objects, they can be moved, changed and otherwise manipulated resulting in a very flexible usage in order to create dynamic scenarios. The training phase is directly influenced by the planning which can be managed in a collaborative way by means of *Photon* networking. This process is based on the action the user takes throughout the briefing and is transferred to the training scene after being accepted by all participants. The concept based on the modular setup of game objects and scripts is easily transferable to various applications and scenarios in different fields of training.

4.2 Managing Collaboration with Custom Roles and Rights

In order to create a dynamic scenario which can be used by various participants in different locations a collaborative room is required. Collaborative in this case means that more than one person can be in a virtual room, recognize other users and interact with them or perform interactions that are made visible to others. Therefore, it is important that each person is in the same virtual room independent of their real-world location. From a technical point of view each user has a local copy and changes, interactions as well as the other relevant information are sent to a specific interest group or all users, after which the data has to be handled. The main room is hosted by a so called master on whose side all settings and interactions in the scene are managed, independent on this user's role. If the master leaves the room before ending the scenario the environment is reset and all information in the scene together with objects created at runtime is lost. In some cases, partial information is sent to another user who becomes the new master. This solution entails various problems for the proposed concept because users take on different roles and each role has its own responsibilities, interaction options, knowledge and rights based on its real-world counterpart (e.g. team leader). In order to prevent a user either obtaining more rights than his role allows or losing objects and interaction possibilities, roles and rights need to be thoroughly managed. As shown in Sect. 4.1 the solution is based on *Unity3D* and the *Photon* SDK automatically creating a virtual master-client room and managing all users in a flat hierarchy. While the master can manipulate the networked room or environment in various ways, other clients can do so only within their local copy, preventing this information from being distributed via the network. If a user is not the master and wants to interact with an object, the existing solution has to be expanded upon. To allow an adaptive, intuitive and collaborative approach that represents an appropriate rights distribution of a real-world scenario like a relationship between teacher and student or organization leaders and employees, this paper defines a custom roles and rights management.

All different relationships of various application areas are reflected in a three-part hierarchy. A group called 'organization' was created which can have different rights and interaction options while remaining on the same hierarchy level. In this concept any number of organizations may be created. An organization consists of one leader and a certain amount of members. Within an organization, the leader has more tools and possibilities at his disposal than a member. The last hierarchy level is the 'own' which does not influence other users and as such reflects the handling of a local copy. This hierarchy concept is shown in Fig. 2 and attached to relevant objects via a script which enables the modular approach described above. Depending on the hierarchy level, users can add objects to the scene or manipulate existing ones. Afterwards, all relevant information such as the object's owner and the associated organization has to be transmitted via the network. The information is automatically sent to all users so that it can be checked who may interact with an object. If the transaction is valid, all connected clients receive an update view. The roles and rights management is based on real-world planning and training of emergency response and civil protection teams [16]. Their expertise was also incorporated in the modular and flexible structure which creates opportunities to adapt the concept to a user-defined scenario. In order to achieve this, the corresponding script automatically transfers all relevant information over the network via *Photon* and manages all important aspects regarding the usability for other users and organizations. Furthermore, it allows for easy extension to include a multitude of organizations as well as adjusting the associated hierarchy. It is also possible to send the information only to a specific group of participants that actually require it, reducing the data volume to a minimum. This information attached to objects can be displayed via UI elements or result in further action such as changing the colour depending on the ownership to visualize the roles and rights in play.

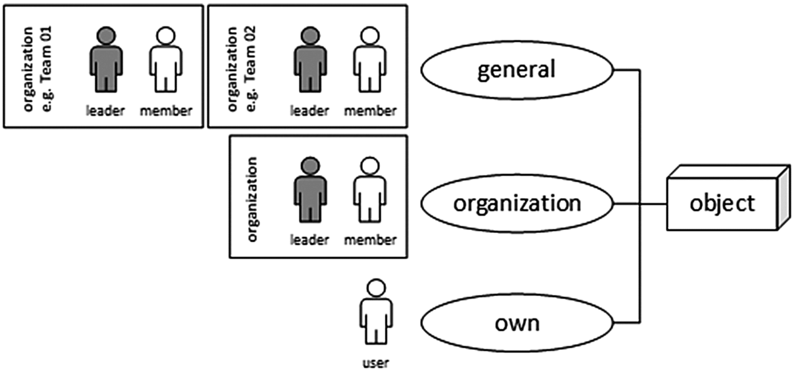


Fig. 2. The designed roles and rights management for object interaction

As part of the roles and rights management an adjustable and automated request handling is integrated that represents real-world interactions required for planning and training. In such a scenario it has to be possible to use or manipulate objects, interact with them or change their state. 'Using' an object means to interact with its functionality, e.g. to open a door, whereas 'manipulating' means to alter the attributes of the object, like changing the size of the door. The custom management defines how users of a certain role may interact with an object. Since there are too many cases to manage individually this is done via an interface. Implemented by the roles and rights management it handles the underlying request necessary for user and object interaction.

The easiest case is a 'general object' like a door. This object is usable by all participants and is not limited to an organization or a specific role. The manipulation of a general object can either be turned on or off entirely for all users and is set when creating the object before the scenario begins. To start an interaction with an object the user only has to touch it with his VR controller. This initiates the 'control' component of the rights management and checks the available options depending on the user. For example, when touching an ‘organization object’ it is determined through the role and the membership of the user whether or not he is allowed to use or manipulate the object. If it is not an object of the same organization the user belongs to, a request is sent to the leader of the owning organization. If the owning organization however is the same the right management feature checks the role of said user. In case of it being the leader the right to use and manipulate the object is automatically granted. If instead the participant is on a lower hierarchy level, he may use the object but only manipulate it after sending a request to the leader (which of course has to be accepted). Lastly, if it is an ‘own object’ the user has the full rights to manipulate and use it. If another actor wants to use such an object, a request is sent to the owner while organizational aspects are omitted. The interaction options depending on the role the user has chosen and the type of object is shown in the following Table 1:

Table 1. Object interaction through use and manipulation in the roles and rights management

Role/ Object	**General**		**Organization** *Team 01*		**Organization** *Team 02*		**Own**	
	Use	Manip-ulate	Use	Manip-ulate	Use	Manip-ulate	Use	Manip-ulate
Leader *Team 01*	✓	✓/✕	✓	✓	⇆	⇆	✓	✓
Member *Team 01*	✓	✓/✕	✓	⇆	⇆	✕	✓	✓
Leader *Team 02*	✓	✓/✕	⇆	⇆	✓	✓	✓	✓
Member *Team 02*	✓	✓/✕	⇆	✕	✓	⇆	✓	✓

legend	
✓	allow
✕	deny
⇆	request

The rights management automatically provides control if a participant is allowed to interact by either using or manipulating an object. The type of access granted is determined through organization, role and ownership of the object as well as the user. However, if a user wants to interact with an object of the same organization but does not have sufficient rights or wants to use an object of a different organization entirely, a request is sent to the person (in most cases the owner) responsible for said object. These requests can be answered by the owner in three different ways in order to give other members rights to the requested object. If none is selected, the request is denied after a certain time so as to not overload the network messaging stack. Figure 3 shows the underlying concept, outlining how the rights management handles the requests of the parties involved depending on their role.

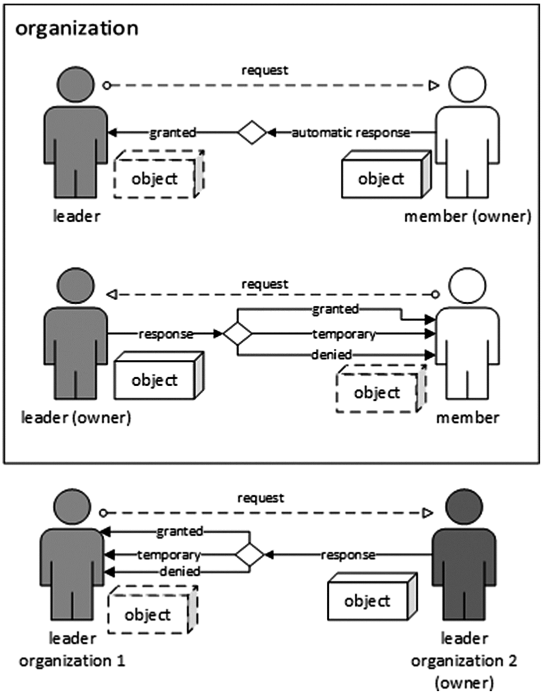


Fig. 3. Request handling in the roles and rights management

The owner has the option to grant the requesting person permission to use the object as if it belongs to them. In this case, the requesting user receives permanent rights until either the owner takes back control or the object is deleted. Another way is to grant temporary access to the object, so that the requesting user can perform precisely one action with the object before the permission is revoked. The last way to answer to a request is to reject it, as a result of which the associated rights remain the same and no permission is granted. Due to the modular approach of handling objects via the roles and rights management the corresponding scripts and game objects in *Unity3D* may easily be adjusted to implement new features. It allows for an intuitive and realistic way of dealing with objects in a given planning and training scenario. On one side the automated request handling and roles management minimizes the necessary interaction and data to handle object rights. On the other side the rights management provides the opportunity to interact with other users in the scenario individually and practice real-world dynamics and relationships, e.g. when an organization leader has to handle his team members as well as the communication and interaction with other organizations. Therefore, the implemented roles and rights management is a powerful tool to handle the objects in a fast and realistic way and a key feature of the proposed planning and training concept.

The interactions between users and objects as well as the actors themselves that are processed with this roles and rights concept are especially key when dealing with the implications of the planning phase. However, many can still be utilized when shifting towards the training phase. By handling all relevant objects in the way described above it becomes possible to create dynamic scenarios with a multitude of different tasks or ways to carry out the designated training routines. This will be demonstrated with a specific collaborative virtual environment in the following section.

5 Implementation of the Collaborative Virtual Environment

In order to validate the concept outlined in Sect. 4 together with the requirements defined in the third section, a use case was constructed that incorporates all necessary features and allows extensive testing of all aspects that were developed to facilitate the collaborative virtual environment. Using the game engine *Unity3D* and the multiplayer game backend *Photon* as described in Sect. 4.1 a safety and security scenario was devised which starts with a detailed planning phase. Once this first phase has concluded, the results of this briefing- or meeting-like session are put into action during the training phase. The setting is an emergency response scenario that takes place inside a building and offers various possibilities to practice certain aspects and procedures. This includes but is not limited to extinguishing fires, navigating the building to locate persons or objects or perform other tasks previously specified in the meeting room. All of the described features can be directly influenced during the planning phase as shown in Fig. 4, for example through varying a team's point of entry, the general location of the fire as well as the specific objects that will burn or the building's layout by blocking certain paths with obstacles or locking doors that could previously be opened. Due to the modular setup it is very easy to narrow down or expand the scope of the exercise or even turn specific features on and off entirely. At the same time, the scene geometry is interchangeable as the program logic can be applied to a broad range of conceivable scenarios.

Before this can be done however, some setup is required. Upon connecting to the room, users may select from different organizations that are related to the given emergency and crisis management scenario, e.g. police, fire brigade or other first responders. They may also simply choose to spectate without taking on a particular role. This initial step also serves as a proof of concept for the roles and rights management described in Sect. 4.2. The leader and member shown in Fig. 4 are representative for various constellations of organizations as well as hierarchy levels that are possible with this implementation. For example, each organization's leader may interact with all objects that belong to the team while the members lower in the hierarchy have to request permission. It is of course entirely possible to assign further interactions such as granting access to everyone in the room, only members of a certain organization, only to specific roles (e.g. all team leaders) or revoke permission for everyone but the object's owner. Due to the requirements defined in Sect. 3, different levels of interaction are visualized, for instance through their color, to help users during the planning phase. Additional feedback is provided through the UI, either utilizing individual and context sensitive menus or the information board that is placed on one of the room's walls. Once a user has selected an appropriate role and the selection was verified through the network, he may move

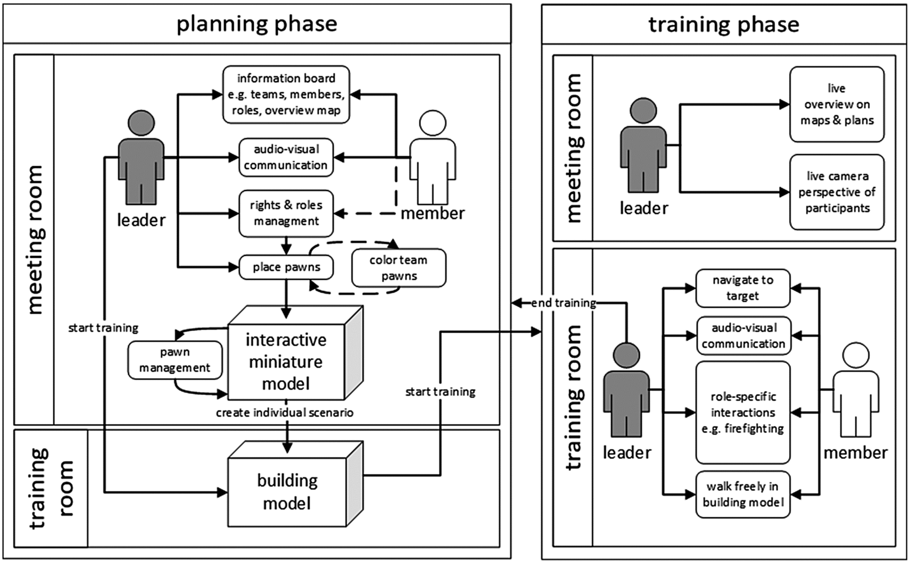

planning phase
meeting room
training room
leader
member
information board e.g. teams, members, roles, overview map
audio-visual communication
rights & roles managment
place pawns
color team pawns
start training
pawn management
interactive miniature model
create individual scenario
building model
start training
training phase
meeting room
leader
live overview on maps & plans
live camera perspective of participants
end training
training room
leader
member
navigate to target
audio-visual communication
role-specific interactions e.g. firefighting
walk freely in building model

Fig. 4. Overview of the collaborative virtual environment

freely inside of the designated area and interact with different objects or communicate with other participants during both phases (s. Fig. 4).

The main feature of the meeting room setup is an interactive miniature model of the building that the training will take place in. Not only does it give all participants an immediate overview of the environment that the training following the planning will be evolved around but it is also used to implement most of the dynamic aspects of that secondary phase based on the results of the first. This feature is also significant to the roles and rights management as a number of users can interact with different objects on varying levels of access. These objects appear in the form of so called pawns as mentioned in Sect. 3 and can be moved around and placed in the interactive miniature model to support the planning phase through visual feedback and provide a unified view of what is happening during the briefing. Additionally, the pawns are scaled to match the size of the model and later used to create a dynamic individual scenario as relevant features are transferred to the building model when the training is started (s. Fig. 4). To exemplify this further, three types of pawns will be discussed in detail. Firstly, each organization's leader may place a marker that represents the spawn point for his team – if it is directly involved in the exercise. While team members can make suggestions by being granted temporary access, the leader has the final say. For now, every team has one spawn point that is symbolized via the figurine-like pawn in the organization's color. On the other hand, location markers can be used to signify important spots, tasks or simply as a waypoint for navigation. All members of an organization can request permission to place one or more of those markers. Lastly, the fire is part of a category of objects that may only be manipulated by a select group of people. As it introduces a whole aspect (in this case fire extinguishing) to the training exercise that (depending on the desired scenario and its learning effects) could be left out entirely and is thus primarily relevant to the fire brigade and their goals this pawn is exclusively handled by those in control of the overall scenario coordination and therefore requires the highest level of access rights.

The setup of the meeting room for the planning phase is completed by a big information board as well as placeholders for different cameras, maps and overviews. In the current version, the information board has a rather generic purpose – it displays the participants currently in the room and any messages or events that are not specific to a certain user or have no place in the personalized UIs attached to the VR controllers. It retains several methods that can be used to display additional information and can fulfill many different purposes in a given scenario or planning session. The wall next to it holds various camera angles and views that can be used to analyze the scenario and function as an important tool that leaders and trainers can use to follow the action and assess trainee performance during the second phase. This could include a live overview on a map or floorplan (s. Fig. 4) that shows key aspects such as a trainee's position or current tasks and location markers. Other possible viewing angles could include live 1st person views from other participants, cameras placed at crucial spots in the building or a map of the surroundings area to plan the approach on a greater scale. All of the features present in the meeting room are once again fully interchangeable and can be swapped in and out depending on the scenario at hand or the goal of a given planning session.

Once briefing and discussion have concluded, the results of the planning phase are carried over to the training phase in order to create and start the dynamic scenario. Only the session leader can initiate the secondary step, at which point all participants that opted to take part in the exercise are moved to the building. As explained before, their starting point is determined by their organization's pawn object that was placed within the interactive miniature model during the planning. Depending on their role, users may immediately start with the appropriate tasks such as search and rescue or firefighting. The latter was used as an example to incorporate real world objects into the training scenario through the *Vive* tracking technology. To accomplish this, a tracker is attached to an actual fire extinguisher that is accurately represented in the virtual environment (s. Figs. 5 and 6). This provides the opportunity for an immersive experience that greatly improves learning effects. The same algorithm that places spawn points throughout the building is also used to transfer other results of planning into training. The tactical approach to a given scenario can be diversified by creating different tasks, blocking or opening specific paths and most importantly changing the general location of the fire as well as placing and moving flammable objects. The manner in which an object burns (and thus the way the fire behaves during the secondary phase) can be influenced via several attributes like heat resistance, ignition temperature and burn time. As such, it is possible to have a big fire spread throughout the building, limit the scenario to certain rooms or even dismiss it in favor of directing the exercise goals entirely towards search and rescue. All of this is done through the interaction with the corresponding objects during the planning phase. While instructors can observe from the planning room using the different tools mentioned above they can also directly join the training room and observe the exercise as shown in Fig. 5.

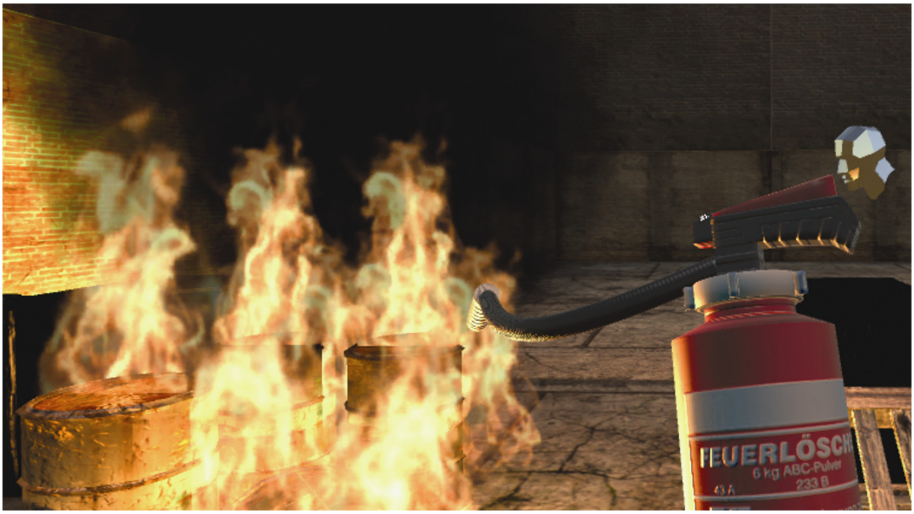


Fig. 5. Collaborative 1st person firefighting training and observation by an instructor (top right)

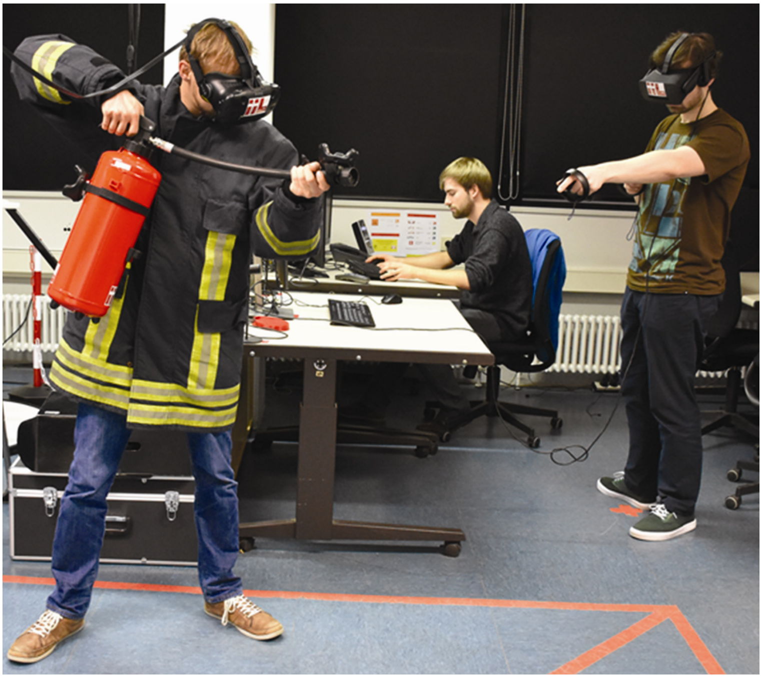


Fig. 6. Multi-user setup from left to right with trainee (*HTC Vive*), spectator (desktop) and instructor (*Oculus Rift*)

Users that are not actively taking part can however still interact with the scenario in general or specific trainees in order to sway the course of events and facilitate the desired learning effects. One example for such an interaction is the navigation feature that guides participants to a certain task or building part and enables real-time influence, e.g. by highlighting a door or room that should be approached next (s. Fig. 4). The corresponding visual components, similar to most other objects, provide intuitive feedback to the users, for example by coloring location markers in accordance to the team that placed them during planning. The objects based on the pawns described above are dynamically placed using *Unity3D*'s resource feature and instantiated through the *Photon* network at runtime. Through the flexible and modular system mentioned in Sects. 3 and 4 it is possible to adjust the level of interaction and thus further individualize the training scenarios. It also incorporates the roles and rights management explained in Sect. 4.2 to prevent invalid or confusing actions where users grab the wrong object or are assigned tasks of another team. Those role-specific interactions are exemplified by fire extinguishers being limited to members of the fire brigade as well as participants only being able to navigate to spots or tasks that where assigned to them during the planning phase. All users can interact with the designated objects during the entirety of planning and training, communicate with one another and move freely within the collaborative virtual environment. The described scenario in an emergency response setting thus showcased the validity of the concept by allowing users with different application types and VR technologies (s. Fig. 6) to learn from planning and training by means of the tasks associated with it. This includes, but is not limited to, the capacity for teamwork by working in a multi-user scenario, improving muscle memory through means of immersive 1st person interaction or practicing situations that might otherwise be difficult or even impossible to carry out. This further reinforces the great potential of such a collaborative virtual environment.

6 Conclusions and Outlook

The approach outlined in this paper was successfully demonstrated using the scenario described in Sect. 5. The fully immersive collaborative environment showed great potential when coupled with a VR application. The designed system utilizes current technologies including HMDs and controller and tracking hardware as well as modern engines for visualization, interaction and networking to establish a dynamic 1st person scenario. This allows users to practice different abilities in a way close to the real world without requiring too much space, having to disrupt daily routines by shutting down buildings or being too costly. It is even possible to go beyond the limits of the real world and tailor a given scenario towards specific needs that could otherwise not be practiced, e.g. by repeating or resetting certain steps. Due to the online networking and the flexible, modular approach the system is mostly independent of physical location or available hardware. The incorporation of real world objects like the fire extinguisher allows participants to learn new interactions that facilitate muscle memory and knowledge alike. As the dynamic training scenario changes every time based on what users came up with during the planning phase this learning effect is reinforced as it is being applied in various ways and settings.

The implementation described here aims to estimate the concept's suitability rather than assessing individual parts of the scenario building or the planning and training in order to determine a final or optimal solution. To accomplish that, existing solutions for certain aspects of a collaborative virtual environment like networking via *Photon* and VR hardware integration through *VRTK* where coupled with highly customizable scripts in *Unity3D*, e.g. the roles and rights management mentioned in the previous sections. The presented concept seamlessly integrates new and existing features making the proposed solution capable of utilizing any building model, various hierarchy levels as well as a number of different interactions with objects and other participants alike. This was achieved through a prototypical implementation which is then adjusted to ensure a suitable course of events and fit one's need regarding the desired planning and training exercise. Here, the authors combined the existing aspects with new features like the roles and rights management, networking relevant parts of the environment and ensuring proper 1st person interactions in VR resulted in an application with a broad scope and the potential to create dynamic scenarios, where specific use cases can utilize all or at the very least a subset of those custom features. The interactive miniature model for example was built from the ground up and tailored specifically to integrate all relevant aspects of planning that are then transferred to the ensuing training, all the while taking roles and rights, collaboration and different user interactions into account.

The emergency response scenario demonstrated the validity of such an approach with a successful implementation of the requirements as well as the concept and design. The exercise as described in Sect. 5 was tested by numerous people including plant fire brigade members of a big German manufacturer. Albeit the demonstrator being limited in its range of features it was well received with the multi-user planning and training as well as the integration of the real-world fire extinguisher being described as very intuitive and thus useful. The immersive interactions facilitate positive effects like learning to navigate a building under ever-changing conditions which further reinforces the need for dynamic scenarios. However, it was also determined that especially users with little to no experience in VR became quickly overwhelmed even though some of functionality was yet to be fully implemented. Therefore, it became apparent that it might be advisable to reduce the extensiveness for specific future applications and utilize the generic approach to build a collaborative virtual environment that focuses on fewer aspects at a time. Due to the flexible and modular setup all of the necessary systems can be adjusted to fulfill specific needs.

In order to easily accomplish that task it would be conceivable to include an accessible scenario builder for certain applications in the future. This way, all of the required functionality of the generic systems like roles and rights management might be incorporated in planning and training of any kind while parts specific to an exercise like the environment or real-world objects could be embedded through the scenario builder. An already possible first step is loading most of the relevant scene geometry from a building information model and afterwards adding components like the network infrastructure or roles and rights that already proved successful. The editor would allow to simply adjust the necessary settings and develop and consequently practice dynamic planning and training. Using the emergency response scenario as an example again, any office building in the world could quickly be made into a planning and training exercise that allows employees to practice things like extinguishing fires or evacuation routines without having to shut down parts of the building or compromising workplace safety. The planning phase would focus on education turning the meeting room into more of a classroom followed by opportunity for a "hands-on" approach. Instructors would be able to connect from anywhere, thus enabling such scenarios to be quickly deployed without the need for extensive on-site personnel.

Before that is possible however, some improvements are required in regards to certain VR modules. Movement and teleportation for example were not focused on during this case study. The teleporter currently retains unlimited range and speed which disrupts immersion and can at times make it difficult to follow other users within the virtual environment. Depending on the situation, the user avatars (heads and hands) can be hard to spot, e.g. behind burning objects or through thick smoke. Possibilities to overcome these problems include "freezing" users for a certain amount of time after they teleport based on the distance covered as well as introducing full body avatars or at least networking more parts that could provide a (realistic) benefit, for example spotting a person's legs or torso during a search and rescue mission. It is also likely that given scenarios might have need for additional tools to put together a proper exercise. In this instance, an interface that seamlessly integrates new aspects into the virtual environment, the networking component as well as the roles and rights management for interactions will prove useful. While the network held up well during testing, integrating more users, objects and interactions might also entail the need for an improved interest management. Another future aim is to increase the number of features available for evaluation and debriefing. In the current implementation, the scenario concludes after the secondary training phase and provides feedback to neither trainees nor instructors. Prospective tools include but are not limited to tracking measurable indicators like time taken for a certain task, recording voice communication and making tasks or entire scenarios repeatable. This would enable instructors to provide detailed feedback, demonstrate better ways of dealing with a situation and generally help with assessing the learning effectiveness of planning and training within a given collaborative virtual environment.

References

1. Redaktion: Südbrandenburger Rettungskräfte proben Ernstfall in Cottbus. http://www.niederlausitz-aktuell.de/cottbus/68807/suedbrandenburger-rettungskraefte-proben-ernstfall-in-cottbus.html
2. Dörner, R., Broll, W., Grimm, P.F., Jung, B. (eds.): Virtual und Augmented Reality (VR/AR): Grundlagen und Methoden der Virtuellen und Augmentierten Realität. Springer, Heidelberg (2013)
3. Guillén-Nieto, V., Aleson-Carbonell, M.: Serious games and learning effectiveness: the case of it's a deal! Comput. Educ. **58**, 435–448 (2012)
4. Merchant, Z., Goetz, E.T., Cifuentes, L., Keeney-Kennicutt, W., Davis, T.J.: Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: a meta-analysis. Comput. Educ. **70**, 29–40 (2014)

5. Monahan, T., McArdle, G., Bertolotto, M.: Virtual reality for collaborative e-learning. Comput. Educ. **50**, 1339–1353 (2008)
6. Greenwald, S.W., Kulik, A., Kunert, A., Beck, S., Fröhlich, B., Cobb, S., Parsons, S., Newbutt, N., Gouveia, C., Cook, C., Snyder, A., Payne, S., Holland, J., Buessing, S., Fields, G., Corning, W., Lee, V., Xia, L., Maes, P.: Technology and Applications for Collaborative Learning in Virtual Reality (2017)
7. Lehner, V.D., DeFanti, T.A.: Distributed virtual reality: supporting remote collaboration in vehicle design. IEEE Comput. Graph. Appl. **17**, 13–17 (1997)
8. Wu, T.-H., Wu, F., Liang, C.-J., Li, Y.-F., Tseng, C.-M., Kang, S.-C.: A virtual reality tool for training in global engineering collaboration. Univ. Access Inf. Soc. 1–13 (2017). https://doi.org/10.1007/s10209-017-0594-0
9. Freeman, K.M., Thompson, S.F., Allely, E.B., Sobel, A.L., Stansfield, S.A., Pugh, W.M.: A virtual reality patient simulation system for teaching emergency response skills to U.S. navy medical providers. Prehosp. Disaster Med. **16**, 3–8 (2001)
10. Pucher, P.H., Batrick, N., Taylor, D., Chaudery, M., Cohen, D., Darzi, A.: Virtual-world hospital simulation for real-world disaster response: design and validation of a virtual reality simulator for mass casualty incident management. J. Trauma Acute Care Surg. **77**, 315 (2014)
11. Sharma, S., Jerripothula, S., Mackey, S., Soumare, O.: Immersive virtual reality environment of a subway evacuation on a cloud for disaster preparedness and response training. In: 2014 IEEE Symposium on Computational Intelligence for Human-Like Intelligence (CIHLI), pp. 1–6 (2014)
12. Passos, C., da Silva, M.H., Mol, A.C.A., Carvalho, P.V.R.: Design of a collaborative virtual environment for training security agents in big events. Cogn. Technol. Work **19**, 315–328 (2017)
13. Auggmed - The serious game platform. http://www.auggmed-project.eu/
14. XVR Simulation: Virtual Reality training software for safety and security. http://www.xvrsim.com/?t=gb
15. LUDUS - Virtual Reality for training in industry and emergencies. http://www.ludus-vr.com/en/
16. Regener, H., Hackstein, A.: Planspiel als Methode und Medium im taktik-und Führungstraining. Rettungsdienst **35**, 24 (2012)
17. Doeweling, S., Tahiri, T., Sowinski, P., Schmidt, B., Khalilbeigi, M.: Support for collaborative situation analysis and planning in crisis management teams using interactive tabletops. In: Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces, pp. 273–282. ACM, New York (2013)
18. Patig, S., Dibbern, J.: Requirements Engineering. http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/is-management/Systementwicklung/Hauptaktivitaten-der-Systementwicklung/Problemanalyse-/Requirements-Engineering/index.html
19. Seyama, J., Nagayama, R.S.: The uncanny valley: effect of realism on the impression of artificial human faces. Presence: Teleoper. Virtual Environ. **16**, 337–351 (2007)
20. HTC Corporation: VIVE^{TM} | Vive Tracker. https://www.vive.com/eu/vive-tracker/
21. Unity Technologies: Unity. https://unity3d.com
22. Exit Games: Multiplayer Game Development Made Easy | Photon Engine. https://www.photonengine.com/en-us/Photon
23. Number of Players Supported by Photon Cloud? — Photon Engine. https://forum.photonengine.com/discussion/3351/number-of-players-supported-by-photon-cloud
24. Welcome to VRTK · VRTK - Virtual Reality Toolkit. https://vrtoolkit.readme.io/docs
25. Cyberith GmbH: Virtual Reality Locomotion Virtualizer Cyberith. https://www.cyberith.com/

3DFacilities: Annotated 3D Reconstructions of Building Facilities

Thomas Czerniawski(✉) and Fernanda Leite

University of Texas at Austin, Austin, TX 78712, USA
tczernia@utexas.edu

Abstract. Scan-to-BIM is the process of converting 3D reconstructions into building information models (BIM). Currently, it involves manual tracing of point clouds by human users in BIM authoring tools, with some automation functionality available for walls, floors, windows, doors, and piping. Emerging semantic segmentation methods demonstrate a level of versatility that could extend the capabilities of automated Scan-to-BIM well past the limited existing object categories. The accuracy of supervised deep learning methods in the context of 3D scene segmentation has experienced rapid improvement over the past year due to the recent availability of large, annotated datasets of indoor spaces. Unfortunately, the semantic object categories in the available datasets do not cover many essential BIM object categories, such as heating, ventilation and air-conditioning (HVAC), and plumbing systems. In an effort to leverage the success of deep learning for Scan-to-BIM, we present 3DFacilities, an annotated dataset of 3D reconstructions of building facilities. The dataset contains over 11,000 individual RGB-D frames comprising 50 scene reconstructions annotated with 3D camera poses and per-vertex and per-pixel annotations. Our dataset is available at https://thomasczerniawski.com/3dfacilities/.

Keywords: Building information modeling · Deep learning · Computer vision

1 Introduction

Building information modeling (BIM) is an increasingly central tool for information management in the architecture, engineering, construction, and facility management industries. It is an attempt to store all relevant building facility information in 3D digital representations called building information models (BIMs), which enable superior information practices as well as new information applications. For an introduction to BIM, see the *BIM handbook* [1]. A comprehensive list of well established and emerging BIM uses was assembled by Change Agents as part of their BIMe initiative[1].

Despite the benefits of using BIM to manage building projects, adoption of BIM has been hampered by the fact BIMs are costly to create and update. This is especially true for the massive collection of existing buildings in the facility operations and maintenance phases of their lifecycles. If an existing building does not already have a BIM, then one needs to be created from scratch, either by the building owner or more

[1] Model Uses List, http://bimexcellence.com/model-uses/, last accessed 2018/01/15.

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 186–200, 2018.
https://doi.org/10.1007/978-3-319-91635-4_10

commonly, an architectural consultant. Once a BIM exists, that model needs to be meticulously updated every time anything changes within the facility due to renovation, remodeling, or maintenance.

In an effort to make BIM adoption easier, researchers have been automating parts of BIM creation by applying reality capture devices and computer vision and 3D modeling algorithms [2–6]. The process begins by digitizing the as-built geometry of existing facilities using sensing technologies such as laser scanners and range cameras. Using this raw spatial data as input, algorithms perform 3D modelling and semantic annotation (Fig. 1). Published work will typically focus on one type of object category such as plumbing [7], partition walls [8], building facades [9], or a limited number of categories together such as walls, floor, doors, and windows [10–14]. Due to the diversity of objects and systems encountered in buildings as well as the diversity of applications and contexts these objects and systems are found in, it is difficult to combine or scale existing methods to fulfill the promise of comprehensive semantically rich BIM creation.

Emerging deep learning methods have demonstrated a level of versatility that could extend the capabilities of automated Scan-to-BIM well past existing methods. Deep learning was acknowledged as a superior computer vision method as a result of the performance of AlexNet [15]. AlexNet is the name of a convolutional neural network that competed in the ImageNet Large Scale Visual Recognition Challenge in 2012 [16]. The competition involved classifying images into 1 of 1000 different possible categories based on the image's content. AlexNet came in first place, achieving an error rate that was 10.8% lower than the second place submission, which was a substantial difference in a competition were other submissions were separated in performance by a few percent, often less than a percent[2]. Research in deep learning applied to computer vision has since extended past 2D images and has experienced success processing 3D data [17–19].

In order to apply these deep learning methods to the process of converting scans of building facilities to building information models, there needs to be available a large, annotated dataset of building facilities. This is because deep learning methods require large datasets of training examples from which to "learn" from. In this paper we present 3DFacilities, an annotated dataset of 3D reconstructions of building facilities (Figs. 2 and 3). The dataset contains over 11,000 individual RGB-D frames comprising 50 scene reconstructions annotated with 3D camera poses and per-vertex and per-pixel annotations. Our dataset is available at https://thomasczerniawski.com/3dfacilities/.

In order to create this dataset, we built a mobile data collection application. The application was run on an iPad with a mounted depth camera. We used the application to collect RGB-D videos that were processed live into 3D reconstructions on the iPad. The 3D reconstructions were then pre-segmented using a 6D DBSCAN based segmentation method [20]. The pre-segmented 3D reconstructions were then refined into semantically segmented 3D reconstructions with instance level category labels using an annotation tool made publically available by Nguyen et al. [21]. We also obtain 2D

[2] Large Scale Visual Recognition Challenge 2012: All results, http://www.image-net.org/challenges/LSVRC/2012/results.html, last accessed 2018/01/15.

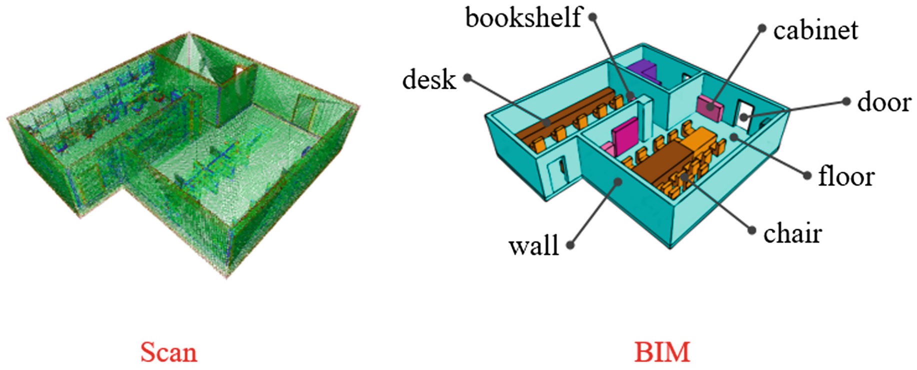


Fig. 1. Scan-to-BIM is the process of converting scans captured by sensing technology into semantically rich building information models.

annotations on the input RGB-D sequences by projecting our 3D annotations into each frame using the corresponding camera pose.

The contributions of this paper are two-fold:

- We present a process for creating datasets that can be used to train deep neural networks for scan-to-BIM; and
- We provide a dataset containing 50 scans that are comprised of RGB-D sequences and their corresponding 3D reconstructions. Each individual RGB-D frame has an associated camera pose as well as a heading, accelerometer, magnetometer, and gyroscope reading. Each individual RGB-D frame and each 3D reconstruction has been hand annotated with instance-level category labels. We hope the availability of this dataset will to promote a data-sharing culture in the Scan-to-BIM community.

2 Related Work

The success of emerging semantic scene segmentation methods is a result of the rapid progress of modern machine learning methods, such as neural models. One requirement for applying these approaches successfully is the availability of large, labeled datasets. Several RGB-D datasets (Table 1) have been created and made available by their authors for training and benchmarking [18, 22–26]. For a comprehensive overview of publicly-accessible RGB-D datasets see [27].

Several commodity RGB-D sensors were involved in the creation of the cited datasets. These include: Kinect v1 [26], Kinect v2 [22, 25], Asus Xtion [22, 25], Intel RealSense [25], Matterport [23, 24] and Structure IO [18].

In order to use these datasets for supervised deep learning, the collected data needs to be both annotated and of sufficient size. This is problematic because the magnitude of the required datasets makes hand labeling individual 2D frames by the authors prohibitively time-consuming, especially in the case of semantic segmentation. In response to this, two alternatives for annotation have emerged. First, authors can outsource the

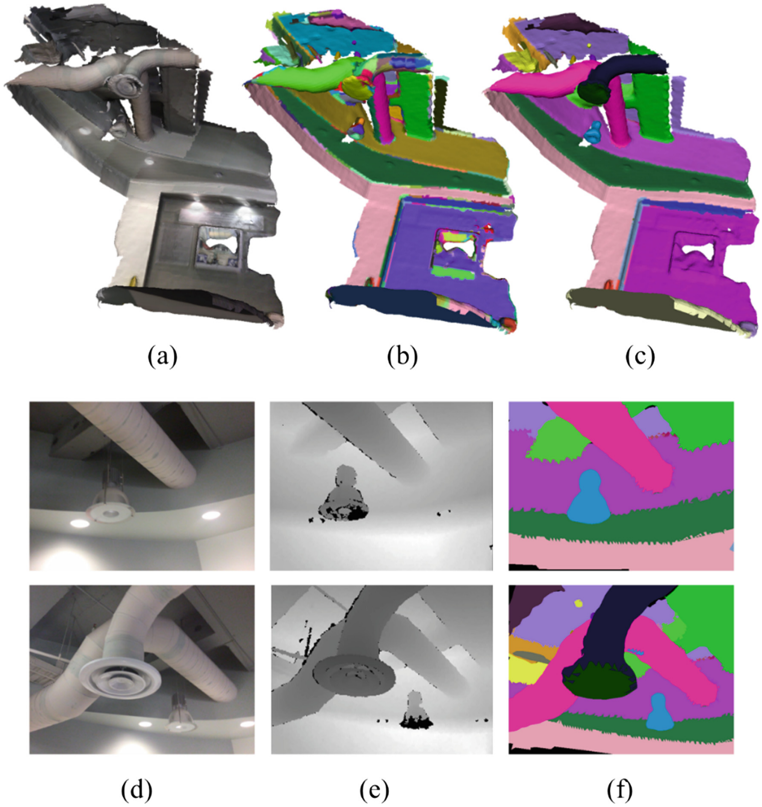


Fig. 2. Example of data provided in the 3DFacilities dataset; (a) 3D reconstruction with color texture (scan); (b) 3D reconstruction pre-segmented using 6D DBSCAN (scan); (c) 3D reconstruction annotated with instance-level category labels (scan); (d) 2D image color (frame); (e) 2D image depth (frame); (f) 2D image annotated with instance-level category labels (frame) (Color figure online)

task of hand labeling through crowdsourcing [25, 26]. This is done by developing an annotation web application and using it in conjunction with a crowdsourcing platform like Amazon Mechanical Turk. On these platforms, crowd workers are assigned annotation tasks and are compensated per task. The second alternative involves generating 3D reconstructions from individual 2D frames, annotating the 3D reconstructions, and projecting the annotations back out to the individual 2D frames [22, 24]. Depending on the number of individual frames used to generate each 3D reconstruction, this process can reduce the number of annotation tasks by several orders of magnitude. ScanNet [18] and Matterport3D [23] combined both methods and annotated 3D reconstructions using crowdsourcing [18, 23].

These datasets have been used to create neural networks that perform many scene understanding tasks including semantic segmentation [17–19] and present a promising tool for Scan-to-BIM technology. However, these large, annotated datasets of indoor spaces, lack many essential BIM categories, such as HVAC and plumbing (Fig. 3).

	3DFacilities (ours)	Matterport 3D [23]	Joint 2D-3D [24]	ScanNet [18]	SceneNN [22]	SUN RGB-D [25]
furniture	🟩	🟩	🟩	🟩	🟩	🟩
door	🟩	🟩	🟩	🟩	🟩	🟩
wall	🟩	🟩	🟩	🟩	🟩	
floor	🟩	🟩	🟩	🟩	🟩	
window	🟩	🟩	🟩	🟩	🟩	
ceiling	🟩	🟩	🟩	🟩		
column	🟩	🟩	🟩			
beam	🟩	🟩	🟩			
stairs	🟩	🟩				
railing	🟩	🟩				
light fixture	🟩	🟩				
elevator	🟩					
plumbing	🟩					
duct	🟩					
diffuser	🟩					
sprinkler	🟩					
cable tray	🟩					
conduit	🟩					
escalator						

Fig. 3. Semantic categories relevant for Scan-to-BIM and the presence of these categories in the annotations of publicly available RGB-D datasets.

Table 1. Overview of RGB-D datasets for semantic scene segmentation. Each of the listed scans and frames are made available by the authors along with associated annotations.

Dataset		Number of scans	Number of frames
[18]	ScanNet	1513	2500k
[22]	SceneNN	100	2500k
[23]	Matterport3D	90[a]	194k
[24]	Joint 2D-3D	–	70k
[25]	SUN RGB-D	–	10k
[26]	NYUv2	–	1k
	3DFacilities (ours)	50	11k

[a] 90 building-scale scenes and 10,900 panoramic views

3 Dataset Acquisition and Annotation Framework

In this section we present the framework used to acquire and annotate the 3DFacilities dataset.

3.1 RGB-D Scanning

Hardware

There are many RGB-D sensors available on the market. In order to be broadly deployable for use in scanning building facilities, we chose our particular camera based on portability, user friendliness, and low-cost. We used the Structure sensor[3]. We attach this sensor to a tablet computer (Fig. 4) — scans in 3DFacilities were collected using an iPad Air 2. The iPad's color camera data is aligned with the depth sensor via a calibration process, providing a synchronized depth and color video. Depth frames are captured at a resolution of 640 × 480 and color at 1296 × 968 pixels.

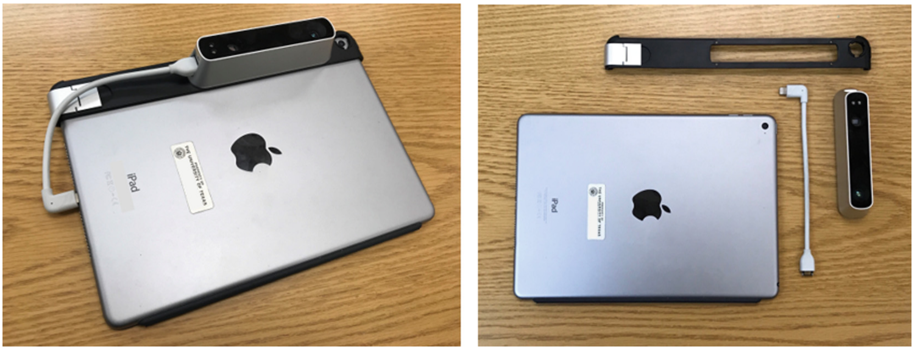


Fig. 4. Structure sensor and iPad Air 2

LivingBIM Data Collection Application

A custom data collection application was created for capturing the 3DFacilities dataset (Fig. 5). The user provides a username and a location for each scan. The location is comprised of both a building name and room number such that when and if the scan is used to update an existing BIM, the scan can more easily be registered to the existing BIM. The user can then proceed to recording a sequence of color and depth frames. The color and depth images are merged live on the iPad into a 3D reconstruction, and the 3D reconstruction is overlaid on the user's view to identify what parts of the scene have been captured and what parts of the scene remain to be captured (Fig. 6). Resulting 3D reconstructions are fairly course as a result of performing reconstruction live on the iPad.

[3] Structure sensor home page, https://structure.io/, last accessed 2018/01/15.

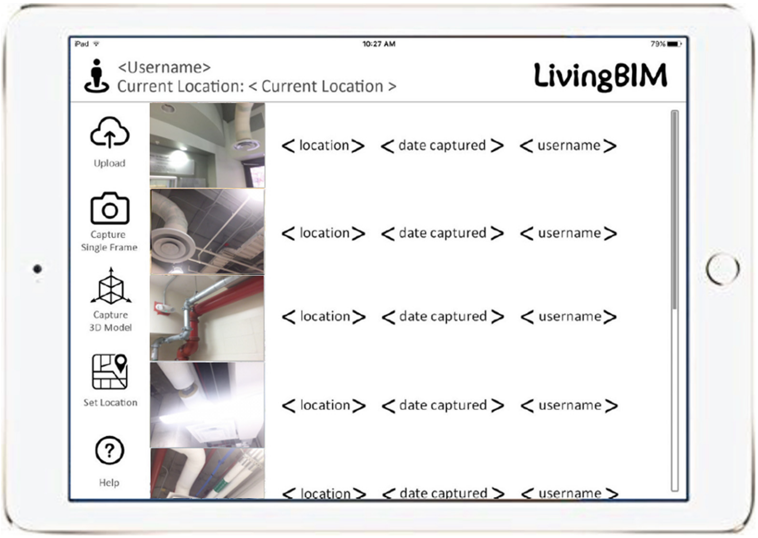

iPad
10:27 AM
79%
<Username>
Current Location: < Current Location >
LivingBIM
Upload
Capture Single Frame
Capture 3D Model
Set Location
Help
< location > < date captured > < username >
< location > < date captured > < username >
< location > < date captured > < username >
< location > < date captured > < username >
< location > < date captured > < username >

Fig. 5. Home screen of the LivingBIM data collection iOS application (Color figure online)

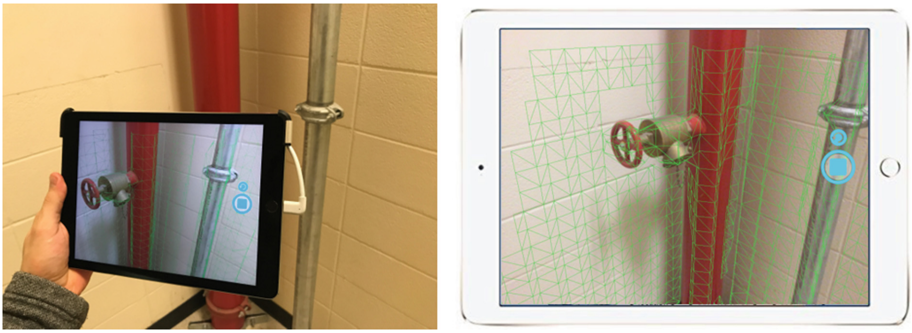


Fig. 6. The LivingBIM data collection iOS application collecting a scan of piping using an iPad and attached Structure sensor. 3D reconstruction is created live and overlaid as a green mesh as individual color and depth frames are collected.

Therefore, after the scan is complete, the 3D reconstructions are refined using a hole filling and smoothing algorithm. The code used to perform reconstruction and hole filling and smoothing is open-source and available at https://structure.io/developers.

We store the individual RGB-D frames at a resolution of 320 × 240 for depth and 640 × 480 for color on the device flash memory. Along with each frame we save a timestamp, Inertial Measurement Unit (IMU) data, and a heading. Once scans have been collected, the user has the option to view the collected scans and upload the scans to cloud storage.

Data Collection
Scans were collected on the University of Texas at Austin campus. Scenes were selected based on accessibility and exposure of building systems.

3.2 Semantic Annotation

We annotated the data by segmenting each 3D reconstruction, assigning each segment a semantic category, and then projecting the annotations of the 3D reconstructions back out to the individual RGB-D frames that comprised the 3D reconstructions. The 3D reconstruction segmentation process had two parts: automatic pre-segmentation and manual annotation.

Automatic Pre-segmentation
First, a 6D DBSCAN based segmentation algorithm was used to automatically pre-segment the scans [20]. The algorithm requires only the x, y, z position of the input mesh vertices along with the surface normal at each vertex. The algorithm provides results similar to curvature based segmentation (Fig. 7) [28].

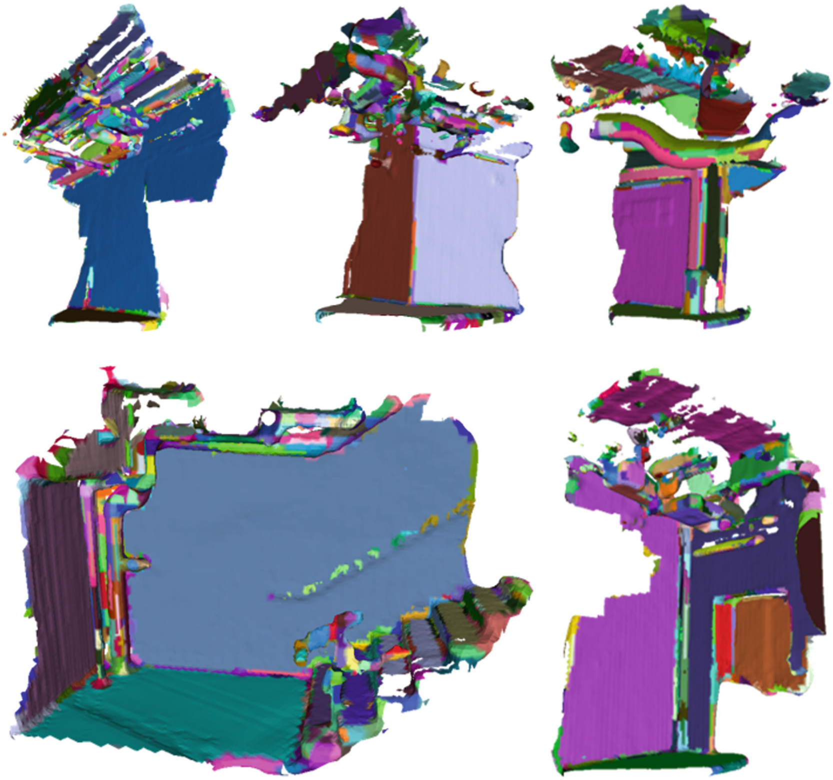


Fig. 7. Example pre-segmented scans

Manual Annotation

The segments were then manually refined by the authors in an annotation tool made publicly available by Nguyen et al. [21]. Each object in the scene was delineated using a segment with a randomly assigned color and then assigned one of 19 category labels: door, wall, floor, window, ceiling, column, beam, stairs, railing, light fixture, elevator, plumbing, duct, diffuser, sprinkler, cable tray, conduit, furniture, or background. Example annotated scans can be seen in Fig. 8.

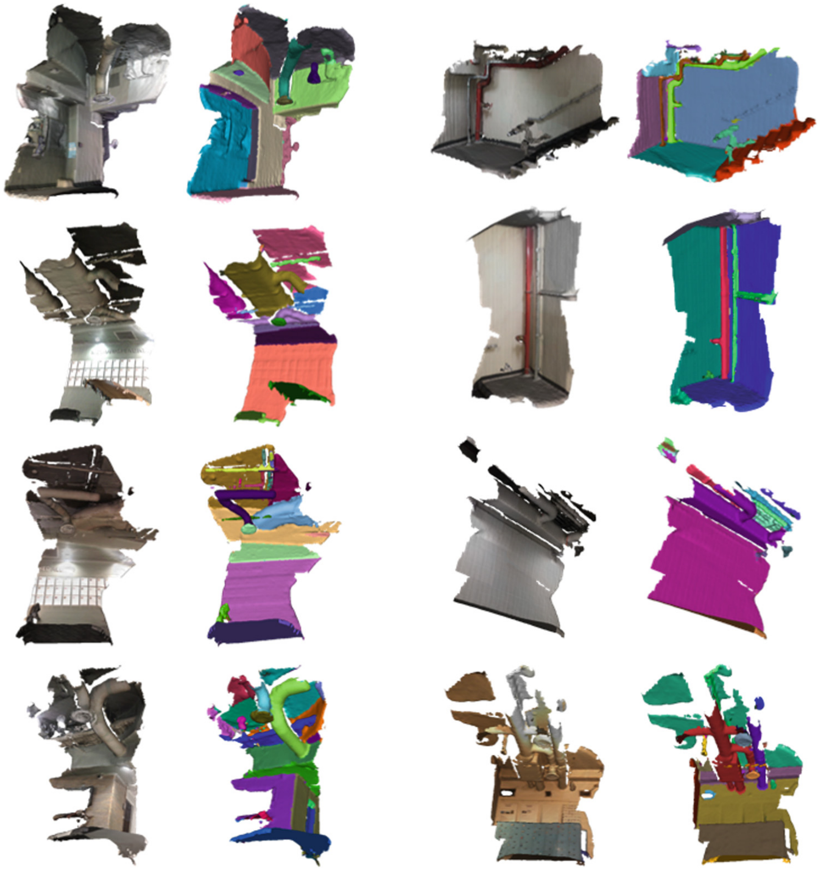


Fig. 8. Example annotated scans in 3DFacilities. **Left:** reconstructed surface mesh with original color texture. **Right:** each object instance shown with a different randomly assigned color. (Color figure online)

The annotations of each 3D reconstruction were then projected onto each individual 2D frame providing 2D annotations on the input RGB-D sequences in addition to the 3D annotations.

The annotations are provided for each scene in an XML file. Each XML files is a series of color and class annotation pairs. For example, one entry in the XML file may be:

```
<annotation>
...
<label id = ″1712960″  color=″255 0 0″  class = 〝duct″>
...
</annotation>
```

Thus, the pixels in the 2D frame and the vertices in the 3D reconstruction with the RGB values of “255 0 0” will have the annotation “duct”. Together, the color and annotation “duct” comprise an instance level annotation.

4 Results and Discussion

4.1 Dataset Statistics

3DFacilities is currently comprised of over 11,000 RGB-D frames and 50 3D reconstructions. Each individual RGB-D frame and each reconstruction has an associated annotation file where each pixel and vertex, respectively, has been categorized into one of 19 different categories. Figure 9 shows the instance count distribution for the 50 reconstructions.

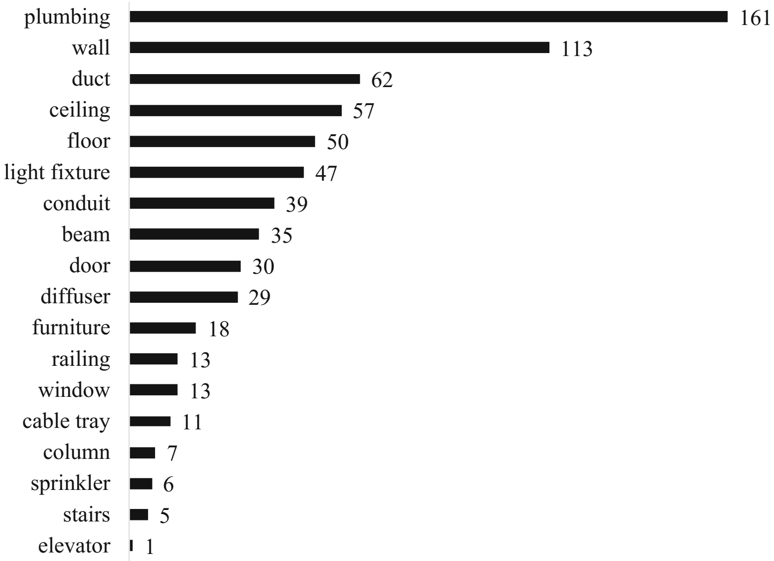


Fig. 9. Instance count for each of the 18 categories in 3DFacilities

4.2 Dataset Validation

The leverage provided by using 3D reconstructions for annotating individual RGB-D frames has drawbacks. The quality of the resulting RGB-D frame annotations is inferior to the annotations that would be provided if each individual RGB-D frame were annotated by hand directly.

In order to quantify this disparity in quality, ten randomly sampled RGB-D frames were annotated directly by hand, providing ground truth annotations, and compared to the

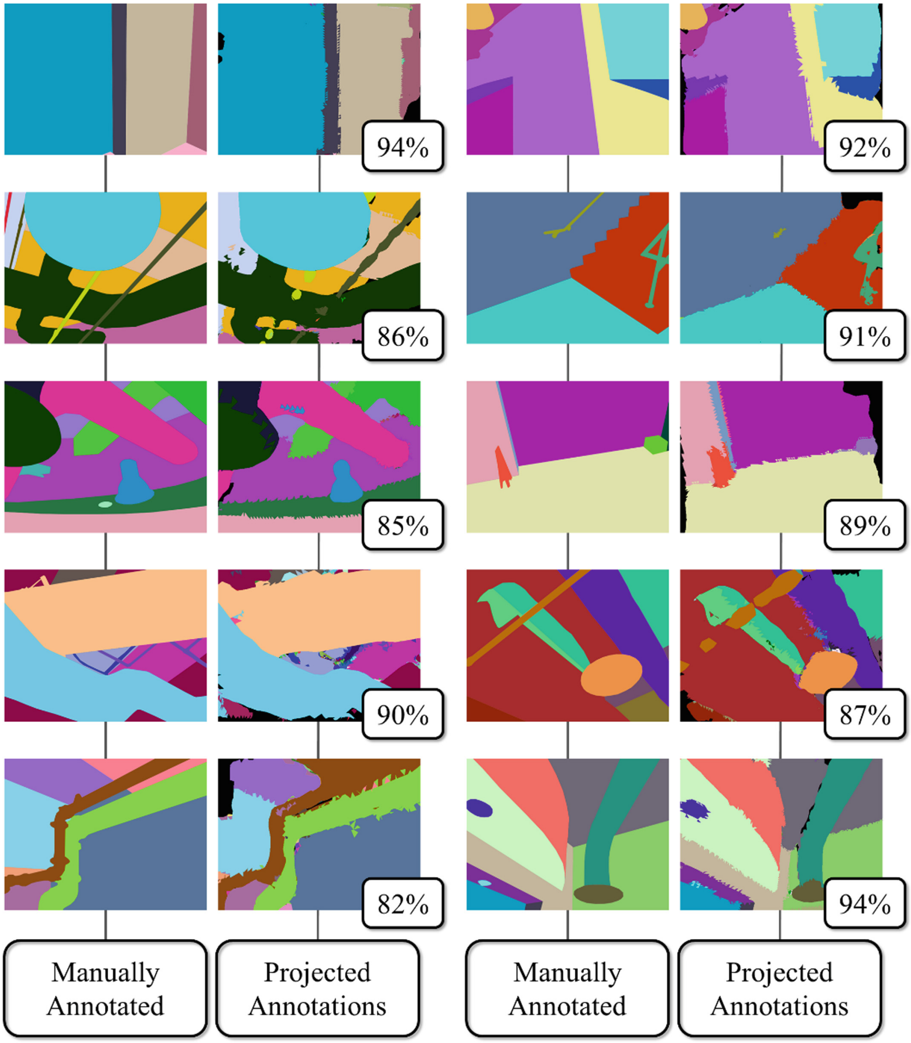


Fig. 10. A comparison of RGB-D frame annotations. On the left of each pair is the manually annotated frame (ground truth) and on the right is the frame annotated using the 3D reconstruction process. Percentage of correctly annotated pixels is indicated for each pair.

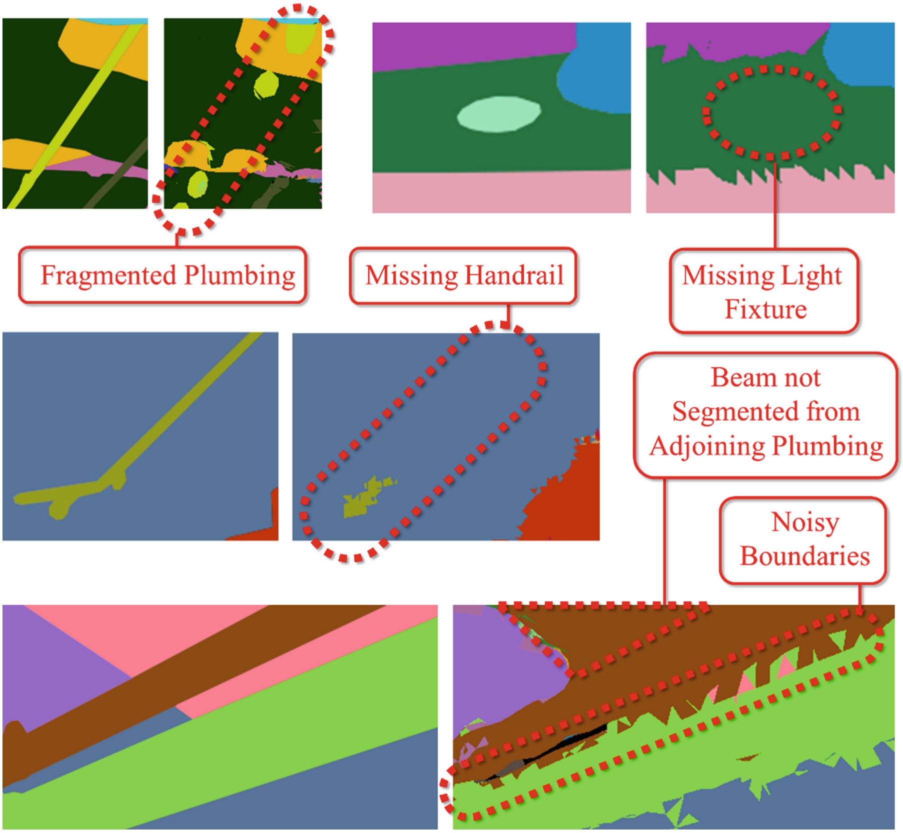


Fig. 11. Examples of issues causing incorrectly annotated pixels of the RGB-D frames

corresponding RGB-D frames provided by the 3D reconstruction method (Fig. 10). Overlapping pixels with the same annotations were classified as being correctly annotated, and overlapping pixels with different annotations were classified as being incorrectly annotated. 89% of the pixels in the RGB-D frames annotated using the 3D reconstruction method matched the annotations of the hand-labelled RGB-D frames. The percent of correctly annotated pixels in each of the ten RGB-D frames can be seen in Fig. 10. Examples of issues that cause incorrectly annotated pixels can be seen in Fig. 11.

4.3 Acquisition Challenges and Dataset Limitations

Collecting scans of building systems is made difficult by the prevalence of metallic objects because current 3D sensing technologies fail to capably detect depth information for shiny, reflective surfaces. This limits collection of data to systems which have been painted over or wrapped in insulation.

Further, when collecting data on a building scale, it can be difficult to detect small components such as small diameter conduit and valves. Many of these fine grained details are lost using the acquisition system presented in this paper. This is a limitation of the hardware as well as the 3D reconstruction process used by the LivingBIM iOS application. In order to perform the reconstruction live on the iPad, a computationally intensive process, the resulting reconstruction is necessarily course. In order to obtain 3D reconstructions with finer detail, an offline reconstruction algorithm would need to be implemented. This has been done by other researchers and is a possible avenue for improving the presented acquisition system [29].

The limitations of the reconstruction process
3DFacilities is noticeably smaller than many of the other publicly available RGB-D datasets. Ideally, a dataset would be as large as possible in order to enable optimal performance by the neural network. 3DFacilities is a specialty dataset, and the authors do not intend the dataset to be used to train neural networks from scratch. Neural networks should be first trained using the larger RGB-D datasets available, and then through transfer learning fine-tuned on 3DFacilities. Regardless, data collection is an ongoing process and 3DFacilities will continue to grow over the coming year.

4.4 Future Work

Scan-to-BIM is in a unique position in the RGB-D community as many of the facilities of interest have existing digital representations in the form of BIMs. Future work for annotating collected data includes transferring building system category labels from BIMs to collected RGB-D frames and reconstructions.

5 Conclusion

This paper introduces 3DFacilities, an RGB-D dataset of over 11,000 frames, 50 reconstructions, and instance-level object category annotations. To make the collection of this data possible, we designed an RGB-D acquisition and semantic annotation framework. The dataset is made publicly available in an effort to promote deep learning enabled semantic segmentation as a foundation for scan-to-BIM development.

Acknowledgments. This research was supported, in part, by the National Science Foundation (NSF) under award number 1562438. Their support is gratefully acknowledged. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Mention of trade names in this article does not imply endorsement by the University of Texas at Austin or NSF.

References

1. Eastman, C., Teicholz, P., Sacks, R., Liston, K.: BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors, 2nd edn. Wiley, Hoboken (2011)
2. Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A.: Automatic reconstruction of as-built building information models from laser-scanned point clouds: a review of related techniques. Autom. Constr. **19**(7), 829–843 (2010). https://doi.org/10.1016/j.autcon.2010.06.007
3. Xiong, X., Adan, A., Akinci, B., Huber, D.: Automatic creation of semantically rich 3D building models from laser scanner data. Autom. Constr. **31**, 325–337 (2013). https://doi.org/10.1016/j.autcon.2012.10.006
4. Volk, R., Stengel, J., Schultmann, F.: Building Information Modeling (BIM) for existing buildings — literature review and future needs. Autom. Constr. **38**, 109–127 (2014). https://doi.org/10.1016/j.autcon.2013.10.023
5. Pătrăucean, V., Armeni, I., Nahangi, M., Yeung, J., Brilakis, I., Haas, C.: State of research in automatic as-built modelling. Adv. Eng. Inf. **29**(2), 162–171 (2015). https://doi.org/10.1016/j.aei.2015.01.001
6. Fathi, H., Dai, F., Lourakis, M.: Automated as-built 3D reconstruction of civil infrastructure using computer vision: achievements, opportunities, and challenges. Adv. Eng. Inform. **29** (2), 149–161 (2015). https://doi.org/10.1016/j.aei.2015.01.012
7. Ahmed, M.F., Haas, C.T., Haas, R.: Automatic detection of cylindrical objects in built facilities. J. Comput. Civ. Eng. **28**(3), 04014009 (2014). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000329
8. Hamledari, H., McCabe, B., Davari, S.: Automated computer vision-based detection of components of under-construction indoor partitions. Autom. Constr. **74**, 78–94 (2017). https://doi.org/10.1016/j.autcon.2016.11.009
9. Oskouie, P., Becerik-Gerber, B., Soibelman, L.: Automated recognition of building façades for creation of As-Is Mock-Up 3D models. J. Comput. Civ. Eng. **31**(6), 04017059 (2017). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000711
10. Quijano, A., Prieto, F.: 3D Semantic modeling of indoor environments based on point clouds and contextual relationships. Ingeniería **21**(3) (2016)
11. Bassier, M., Vergauwen, M., Van Genechten, B.: Automated semantic labelling of 3D vector models for scan-to-BIM. In: Proceedings of the 4th Annual International Conference on Architecture and Civil Engineering (ACE2016), pp. 93–100 (2016)
12. Ochmann, S., Vock, R., Wessel, R., Klein, R.: Automatic reconstruction of parametric building models from indoor point clouds. Comput. Graph. **54**, 94–103 (2016). https://doi.org/10.1016/j.cag.2015.07.008
13. Anagnostopoulos, I., Pătrăucean, V., Brilakis, I., Vela, P.: Detection of walls, floors, and ceilings in point cloud data. In: Construction Research Congress, pp. 2302–2311 (2016). https://doi.org/10.1061/9780784479827.229
14. Mura, C., Mattausch, O., Pajarola, R.: Piecewise-planar reconstruction of multi-room interiors with arbitrary wall arrangements. Comput. Graph. Forum **35**, 179–188 (2016). https://doi.org/10.1111/cgf.13015
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
16. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, A., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)

17. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3d classification and segmentation. arXiv preprint arXiv:1612.00593 (2016)
18. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: richly-annotated 3d reconstructions of indoor scenes. arXiv preprint arXiv:1702.04405 (2017)
19. Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and multi-view CNNs for object classification on 3d data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5648–5656 (2016)
20. Czerniawski, T., Sankaran, B., Nahangi, M., Haas, C., Leite, F.: 6D DBSCAN-based segmentation of building point clouds for planar object classification. Autom. Constr. **88**, 44–58 (2018). https://doi.org/10.1016/j.autcon.2017.12.029
21. Nguyen, D. T., Hua, B.S., Yu, L.F., Yeung, S.K.: A robust 3d-2d interactive tool for scene segmentation and annotation. arXiv preprint arXiv:1610.05883 (2016)
22. Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K.: SceneNN: a scene meshes dataset with annotations. In: The IEEE Fourth International Conference on 3D Vision (3DV), pp. 92–101 (2016)
23. Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., et al.: Matterport3D: learning from RGB-D data in indoor environments. The Computing Research Repository (CoRR), 1709.06158 (2017)
24. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2D-3D-semantic data for indoor scene understanding. The Computing Research Repository (CoRR), 1702.01105 (2017)
25. Song, S., Lichtenberg, S.P., Xiao, J.: Sun RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 567 (2015)
26. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Computer Vision–ECCV, pp. 746–760 (2012)
27. Firman, M.: RGBD datasets: past, present and future. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 19–31 (2016)
28. Dimitrov, A., Golparvar-Fard, M.: Segmentation of building point cloud models including detailed architectural/structural features and MEP systems. Autom. Constr. **51**, 32–45 (2015). https://doi.org/10.1016/j.autcon.2014.12.015
29. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. ACM Trans. Graph. (TOG) **36**(3), 24 (2017)

Meta Models for Real-Time Design Assessment Within an Integrated Information and Numerical Modelling Framework

Jelena Ninic[1(✉)], Christian Koch[2], and Walid Tizani[1]

[1] University of Nottingham, Nottingham NG7 2RD, UK
jelena.ninic@nottingham.ac.uk
[2] Bauhaus-Universität Weimar, 99423 Weimar, Germany

Abstract. In situations where rapid decisions are required or a large number of design alternatives is to be explored, numerical predictions of construction processes have to be performed in near real-time. For the design assessment of complex engineering problems such as mechanised tunnelling, simple numerical and analytical models are not able to reproduce all complex 3D interactions. To overcome this problem, in this paper a novel concept for on-demand design assessment for mechanized tunnelling using simulation-based meta models is proposed. This concept includes: (i) the generation of enhanced simulation-based meta models; (ii) real-time meta model-based design assessment in the design tool, and; (iii) the implementation within a unified numerical and information modelling platform called SATBIM. The capabilities of this concept are demonstrated through an example for the evaluation of tunnel alignment design and the assessment of the impact of tunnelling on existing infrastructure. Moreover, meta models are used for fast forward calculation in sensitivity analyses for the evaluation of the importance of model parameters. The concept proved its efficiency by assessing the design alternatives in real-time with the prediction error of less than 3% compared to complex numerical simulation in presented example.

Keywords: Building Information Modelling · Numerical analysis
Meta models · Level of detail · Soil-structure interaction · Real-time prediction

1 Introduction

Complex engineering problems, such as mechanized tunnelling, require reliable design assessment from early design over to construction and operation phases. In situations where quick decisions are required, or when a large number of alternatives has to be tested, the predictions have to be performed in near real-time. This can be achieved using analytical and empirical solutions. However, these introduce a number of assumptions and simplifications associated with them. For example, analytical and empirical solutions do not take three-dimensional effects and complex interactions between individual components into account, and they are usually characterized with a simple linear elastic response [1]. On the other hand, complex 3D numerical simulations are able to reproduce all complex soil-structure interaction effects induced by the tunnelling process, however, they are often characterised by high computational cost,

© Springer International Publishing AG, part of Springer Nature 2018

I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 201–218, 2018.
https://doi.org/10.1007/978-3-319-91635-4_11

and are difficult to operate in real-time [2]. This can be overcome using computationally efficient meta models instead of the original, detailed numerical models [3].

Meta modelling is understood differently in different domains. In model-driven software engineering, for example, a meta model specifies the structure, the semantics, and the constraints for a family of models in a certain domain, e.g. in cyber-physical systems modelling [4]. While a model is, simply speaking, an abstraction of phenomena in the real world, a meta model is a further abstraction that specifies the properties of the model itself [5], In other domains, meta models have also been developed to serve as "surrogate models" for expensive simulation processes in order to improve the overall computational efficiency. In that sense, they have been applied to solve a number of practical engineering problems in the last years. Meta models are extensively used for prediction, sensitivity analysis, pattern recognition, and design optimization [6]. In tunnelling applications, meta models trained using field monitoring data collected during the tunnel construction process [7] or complex simulation models [8] have been applied for predicting the surface settlements induced by tunnelling. Apart from the prediction speed, the advantage of using meta models is their ability to learn from the different types and large amount data and therefore interpret and summarise existing knowledge in different forms compared to physical models.

Simulation models, on the other hand, are complex and require a large amount of project-specific information that is often available in the form of dispersed resources usually either given in some Computer Aided Design (CAD) format, or, more recently, stored in a Building Information Model (BIM) together with other relevant data about design and construction [9, 10]. One of the challenges during the optimisation of a project design is to preserve the consistency between design alternatives and the corresponding design assessment across different sub-models and throughout different phases. Currently, this is usually an error-prone process, involving manual conversion of data from a BIM or similar storage to a format accepted by numerical design tools. An efficient solution to solve this problem is an integrated design-analysis-assessment framework where the numerical simulations are automatically generated based on the geometry and semantics stored in BIM design tools such as Revit [11]. Therefore, we proposed a unified platform for information, numerical modelling and visualisation of simulation results called "SATBIM" (Simulations for multi-level Analysis of interactions in Tunnelling based on the Building Information Modelling technology) [12]. In this paper, this platform is extended with a tool for meta model-based design assessment.

In the unified design-analysis-visualisation platform SATBIM, we developed a novel concept for on-demand design assessment at the design phase using simulation-based meta models, as "surrogate models", for real-time prediction. To this end, Sect. 2 presents: (i) a brief description of the unified design-analysis-assessment platform; (ii) the concept for on-demand real-time design assessment using simulation-based meta models; (iii) the meta modelling techniques and requirements for generation of enhanced meta models, and (iv) the importance of sensitivity analysis in this concept. The implementation of this concept within the unified numerical and information modelling platform SATBIM is given in Sect. 3. Finally, in Sect. 4, we present a numerical example for the evaluation of tunnel alignment design and the assessment of the impact of tunnelling on existing infrastructure. Moreover, the importance of model parameters is evaluated by means of sensitivity analysis.

2 Methodology

In order to enable on-demand design assessment in engineering design environments, a unified platform for information and numerical modelling considering a multi-level representation, extended with a tool for real-time prediction is proposed in this paper.

2.1 On-Demand Design Assessment in Information Models

Unified Platform for Information and Numerical Modelling. SATBIM is an integrated, open-source platform for information modelling, structural analysis and visualisation of the mechanised tunnelling process. Based on a parametric BIM for tunnelling [13], a multi-level simulation model is developed to support engineering decisions during the project life cycle and to allow for the evaluation and minimisation of risks on existing infrastructure (see Fig. 1). Within this platform, industry-standard tools (Autodesk Revit) are employed for the design of the tunnel structure and the surrounding infrastructure with consideration of different Levels of Detail (LoDs) for all system components (soil, tunnel structure, tunnel boring machine, existing buildings). Based on the multi-level, parametric BIM, multi-level numerical models for each component are developed, considering proper geometric as well as material representation, interfaces and the representation of the construction process. The numerical models are then fully automatically instantiated and executed based on the geometry and semantic exported from BIM using newly developed software SatBimModeller [12]. Finally, the simulation outputs are read back and visualised within Revit.

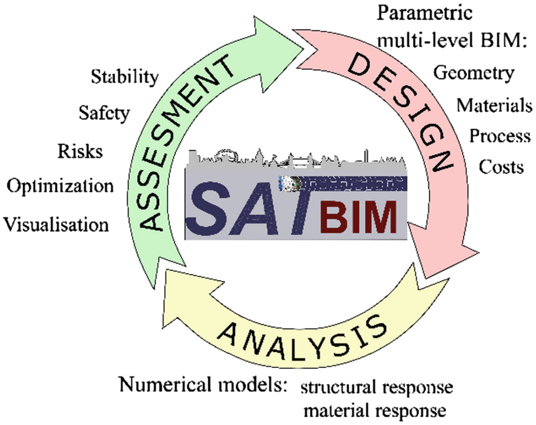


Fig. 1. Concept for integrated SATBIM platform for design, numerical analysis and assessment on different LoDs.

In the SATBIM framework, a multi-scale modelling concept is applied to the shield tunnelling components (soil, TBM, lining, buildings) to enable efficient representation of the tunnelling process with different LoDs as the calling process requires. For example, on the kilometre scale, a low LoD is applied to represent the general alignment of the track and surrounding infrastructure, while on the centimetre scale, all details are captured with the highest LoD of each component. The multi-level approach is also useful over different project stages due to the availability of the information and details at the different design phases. At early design stages only basic requirements are known, and therefore lower LoDs can be represented, while towards the final design and over the construction phase the highest LoDs are represented within information and numerical models. Such an integrated multi-scale design-numerical approach contributes to modelling efficiency by minimising the time needed for model generation as well as computation.

This model is used as a basis for (i) information modelling, (ii) numerical modelling for the generation of the data set for meta model training and (iii) visualisation of numerical assessment results.

Real Time-Design Assessment. To enable real-time assessment, meta models are trained a priori using a process-oriented simulation model generated from a multi-level tunnelling information model (TIM) using the SatBimModeller [12]. Apart from settlements, output parameters include the lining stresses, pore pressures, and damage estimates for existing buildings. Figure 2 illustrates the use of simulation-based meta models for real-time predictions. For different design alternatives, with particular design parameters, simulation models are automatically generated and executed using the SatBimModeller. The simulation results are stored in a format suitable for meta model training. The data set obtained from the simulation model is trained by means of a hybrid training algorithm (described below) to create enhanced meta models. Finally, the resulting meta model is implemented in Dynamo (a visual programming tool for Revit) to enable interactive visualisation of the effects of design choices within the Revit design environment.

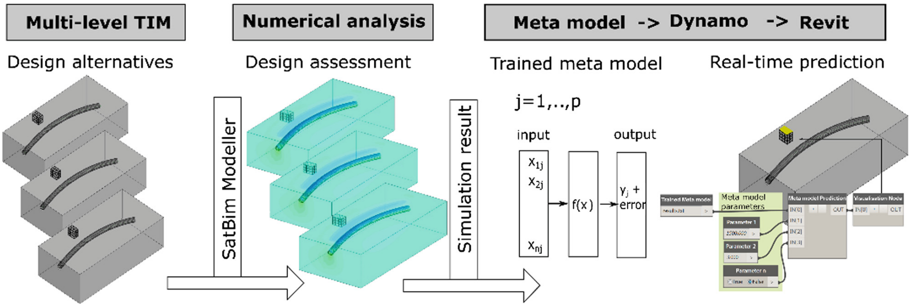


Fig. 2. Workflow of real-time predictions for design assessment within SATBIM.

2.2 Meta Models for Real-Time Prediction

Machine Learning Methods. For the purpose of real-time predictions of tunnelling-induced effects such as surface settlements, risk on building damage, etc., a meta model is employed to substitute for computationally demanding 3D numerical simulations. An algorithm is developed to select an optimal meta model by evaluating and comparing different methods for data training:

- Polynomial Regression (PR),
- Artificial Neural Networks (ANNs),
- Support Vector Regression (SVR) (with Radial Basis Function (RBF) kernel SVR-RBF and Polynomial kernel (SVR-Poly).

In the following, the fundamentals of the prediction models used in this paper are described.

Polynomial Regression. This is a meta modelling approach for modelling the relationship between a scalar dependent variable y (output or target variable) and one or more independent variables **x** (in our case the input vector). For given a data set $\{y_i, , x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of *n* patterns, a polynomial regression model assumes that the relationship between the dependent variable *y* and the *p* vector of input variables x_i is modelled as an *n*th degree polynomial in *x* [14].

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_m x_i^m + \varepsilon_i \; (i = 1, 2, \ldots, n) \tag{1}$$

Where $\boldsymbol{\beta}$ are regression coefficients and an ε_i is an “error variable” that adds noise to the polynomial relationship between the dependent variable **y** and inputs **x**. In our application, we are using second-order polynomials, so that the model now becomes:

$$y_i(\beta, x) = \beta_0 + \sum\nolimits_{i=1}^{n} \beta_i x_i + \sum\nolimits_{i=1}^{n} \beta_{ii} x_i^2 + \sum\nolimits_{i=1}^{n} \sum\nolimits_{j>i}^{n} \beta_{ij} x_i x_j = \boldsymbol{X}^T \boldsymbol{\beta} + \varepsilon_i \tag{2}$$

The vector of estimated polynomial regression coefficients is estimated using the Ordinary Least Squares (OLS) method, as the simplest and thus most common estimation method. The OLS method minimizes the sum of squared errors and residuals in statistics, and leads to a closed-form expression for the estimated value of the unknown parameter β:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{3}$$

This is the unique least-squares solution as long as $\boldsymbol{X}$ has linearly independent columns.

Artificial Neural Networks. This method of machine learning, as an attempt of mimicking the human brain and neural learning, is capable of learning the pattern associated with a large body of data [2, 15]. In this method, the relationship between the input parameters x_i (accepted through the input neurons) and output y_k (corresponding to output neurons) is established through a network of hidden neurons with corresponding connection weights **w**:

$$y_k(x, w) = f\left(\sum\nolimits_{j=1}^{m} (w_{jk} + \theta_j) f \sum\nolimits_{i=1}^{n} (w_{ij} x_i + \theta_i)\right) \tag{4}$$

In this equation, w_{ij} are connection weight between input and hidden neurons, w_{jk} are connection weights between hidden and output neurons, θ is bias and $f()$ an activation function used to transform the input values and transfer them to the next layer. The relation between the input and output is established by a training procedure, adjusting the connection weights $\boldsymbol{w}$ in order to minimise the error E between output $\boldsymbol{y}$ and target values $\boldsymbol{t}$ for number of patterns m. Using a gradient descent method, the connection weights are updated as follows:

$$w_{is} = w_{is-1} + \Delta w = w_{is-1} - \gamma \frac{\partial E}{\partial w} \quad where \quad E = \sum\nolimits_{k=1}^{m} (y_k - t_k)^2 \tag{5}$$

Where γ is a learning rate. The training procedure is performed in a number of iteration steps *is* and the weights $\boldsymbol{w}$ are updated for all connections between input-hidden, hidden-hidden, and hidden-output neurons. The quality of the meta model training depends on both (i) the network architecture (number of hidden layers and hidden neurons) and learning parameters (number of iteration steps and learning rate γ).

Support Vector Regression. This is a machine learning method where the so-called support vectors determine the approximation function [16]. In this method the multivariate regression function *f(x)* is established based on the input data set x to predict the output data $\mathrm{y} = \mathrm{f(x)}$ such as:

$$y = f(x) = \sum\nolimits_{x_i \in SV} (\alpha_i - \alpha_i^*) K(x, x_i) + b \tag{6}$$

where K is a kernel, n is the number of training data, b is an offset parameter of the model, α_i and $\alpha_i^* \neq 0$ are Lagrange multipliers of the primal-dual formulation of the problem, and SV is the support vector set. The transformed regression problem may be solved by quadratic programming and only the input data corresponding to the non-zeros α_i and α_i^* contribute to the final regression model.

The kernel K is a non-linear mapping from an input space onto a characteristic space through a dot product of the non-linear kernel function $\phi(x)$. In this application, two different types of kernel functions are tested:

- Polynomial function: $(\gamma \left(x, x\prime\right) + r)^d$ where d is the polynomial degree and r is an independent coefficient
- Radial basis function $(-\gamma \left\| x - x\prime \right\|^2)$, with γ is a coefficient greater than 0.

Enhanced Meta Model. In all training methods, the simulation model input parameters x_i are taken as input variables for the meta model training while the tunnelling-induced effects, i.e. outputs of complex numerical summations (settlements, risk of damage, etc.) are target values for meta model training. Meta models are then trained to predict the output for given input values minimising the error between the target and output values. In order to achieve the best prediction capabilities of the meta model, the following steps are taken:

- data normalisation,
- data split, and
- optimisation of free parameters of machine learning methods.

In order to achieve better training performance, all input-output pairs are normalised, i.e. mapped to the interval [0.1; 0.9] using a data normalization algorithm. For a parameter V the normalized value V_{norm} is obtained as

$$V_{norm} = \frac{V - V_{min}}{V_{max} - V_{min}} (\bar{V}_{max} - \bar{V}_{min}) + \bar{V}_{min} \tag{7}$$

where V_{max} and V_{min} are the maximal and minimal value of the variable V, and $\bar{V}_{max}$ and $\bar{V}_{min}$ are the maximal and minimal values of the variable V after normalization, defined as 0.1 and 0.9.

For the training of the enhanced meta model, the data set is split into data for training, testing and validation of the meta model, according to prescribed portions, splitting the data of the whole set at random (see Fig. 2). The learning process is performed with the training set, while the test set is used to test the prediction performance. Finally, the meta model quality is evaluated with the validation set. This data split is important (i) to avoid model overfitting and (ii) to double-check the prediction capabilities of the trained meta model.

Some of the mentioned machine learning methods are characterized by having parameters which influence the training performance (e.g. neural networks: number of hidden layers, nodes and learning rate). In order to have enhanced meta models, those parameters are optimized with the Particle Swarm Optimization (PSO) method [17] as shown in Fig. 3.

In the PSO method, the system is initialized with a population of random solutions. PSO then searches for optima by updating subsequent generations. The potential solutions, called particles, “fly” through the problem space by following the current optimum particles. Each particle belongs to a swarm and has two properties: velocity (v_{ij}) and position (x_{ij}). The particle keeps track of its coordinates in the problem space, which are associated with the best solution (fitness), and achieves the particle-best value *pbest* (x_{ij}^{pbest}). If a particle takes the complete population as its topological neighbours, the best value is a global best *gbest* (x_{ij}^{gbest}). The new velocity and position of the particles are updated in each iteration using the following equations:

$$\begin{aligned} v_{i,j+1} &= w_{ij} + \phi_1 r_1 \left(x_{ij}^{pbest} - x_{ij} \right) + \phi_2 r_2 \left(x_{ij}^{gbest} - x_{ij} \right) \\ x_{i,j+1} &= x_{ij} + v_{i,j+1} \end{aligned} \tag{8}$$

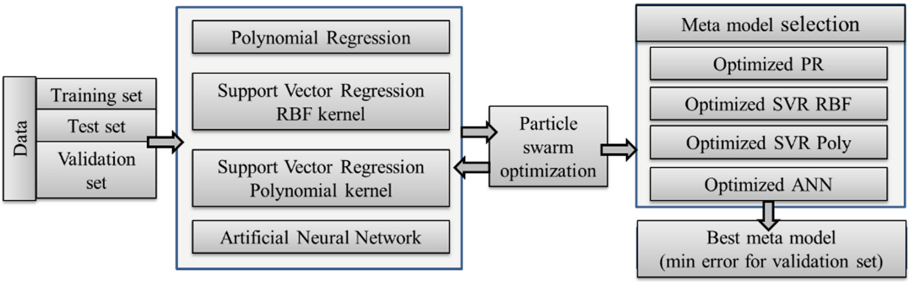

Data
Training set
Test set
Validation set
Polynomial Regression
Support Vector Regression RBF kernel
Support Vector Regression Polynomial kernel
Artificial Neural Network
Particle swarm optimization
Meta model selection
Optimized PR
Optimized SVR RBF
Optimized SVR Poly
Optimized ANN
Best meta model (min error for validation set)

Fig. 3. Algorithm for generation of enhanced meta models. Optimisation of free parameters with the PSO method and selection of the best model with minimum error on the validation set.

Where w are weights, r_1 and r_2 represent random numbers uniformly distributed over [0; 1] and ϕ_1 and ϕ_2 are so-called cognition and social learning factors.

The details about the optimisation algorithm are presented in Sect. 3.1.

2.3 Sensitivity Analysis for Model Evaluation

Sensitivity analysis is a vital tool in the SATBIM framework for performing the following tasks:

- determination of the sensitivity/importance of the component LoD for a defined analysis scenario and w.r.t. design parameters,
- general study of sensitivity/importance of design parameters (geometrical, material and process) for predefined analysis scenarios, and
- generation of reliable meta models based on important parameters determined by the sensitivity analysis.

In this paper, we give an example of how sensitivity analysis can be used to evaluate the importance of input parameters for design assessment. For this purpose, the One at Time (OAT) design method is used. This method is based on the discretization of the inputs in levels, allowing a fast exploration of the model behaviour and identification of the influential inputs. In this variance-based method, the importance of the input parameter is quantified through (i) the absolute mean μ^* of the elementary effect EE_i, representing the total sensitivity index and a measure of the overall effect of a factor on the output and (ii) the standard deviation σ, which detects the interaction effects with the other parameters as well as the nonlinear relation between the corresponding input/output [18, 19]. The elementary effect EE_j^i of the j^th parameter X_j in the i^{th} repetition as well as σ and μ^* are calculated as:

$$EE_j^i = \frac{Y\left(X_1, \ldots X_j + \Delta, \ldots X_k\right) - Y\left(X_1, \ldots X_j, \ldots X_k\right)}{\Delta};$$

$$\sigma_j = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} \left\| EE_j^i - \mu_j \right\|} \quad and \quad \mu_j^* = \frac{1}{n}\sum\nolimits_{i=1}^{n} \left\| EE_j^i \right\| \tag{9}$$

where *Y(X_j)* is an output and Δ is a predetermined multiple of $1/(p-1)$, with p denoting the number of intervals of X_j.

If predictions in real-time are required and the simulation-based meta models are used as a tool, it is necessary to ensure the robustness and reliability of those meta models. One very important aspect of the robustness is that the meta models are defined with input parameters which are denoted as "important", i.e. which have a significant effect on the output. This can be ensured by performing sensitivity analysis a priori to meta model generation. This will be an important matter of further research within SATBIM framework.

3 Workflow and Implementation

In this section, the details of the implementation of robust meta models, visualisation of the simulation results and sensitivity analysis for evaluation of the numerical models are described.

3.1 Implementation of the Enhanced Meta Models

The algorithm for the robust meta model training is implemented in Python following the main idea described in Fig. 3 and applying the methods of machine learning described in Section "Machine learning methods". For the implementation, the Python library SciKitLearn for supervised learning is used [20]. This toolkit contains implementation of regression models, ANNs and SVR. However, in order to achieve robust learning, the PSO method was implemented to optimize training parameters of different machine learning models. Moreover, the following additional methods were implemented:

- *ReadDatasetFile()* for reading the data set with its associated arguments file, and the training portion, test portion, validation portion;
- *ComputeError()* with arguments training method, training set, test set;
- *ForwardPass()* with arguments weights, data set, method parameters;
- *ViewPerformance()* with arguments training set, test set, validation set.

The main function for robust meta model training is given in Listing 1. As outlined in Listing 1, the free parameters of each machine learning methods are iteratively updated with the PSO method. To this end, the free parameters are initialised as particles characterised with position x_{ij} and velocity v_{ij}, moving through the solution space iteratively updating according to Eq. (8) with the objective to minimise the prediction error of the validation set.

```
"Training, testing and validation of meta models."
def main():
    methods = [ linear_model.SGDRegressor(),
        svm.SVR(kernel='rbf', C=5, gamma=0.4),
        svm.SVR(kernel='poly', C=1e3, degree=3),
        MLPRegressor(solver='lbfgs', alpha=1e-4, random_state=1, max_iter=5000]
"Optimize model parameters."
   for clf in methods:
           param_svr_new = PsoOptimize(clf, model_param_, bx, by, bxt, byt, 20, 20 )
           clf.set_params(C = param_svr_new[0])
           clf.fit(bx, by)
           rrmse = ComputeError(clf, bxt, byt)
           tot.append(rrmse)
 "Choose best model, test and plot results."
    clf = meth[BestModel(tot)]
    rrmse_test = ComputeError(clf, bxt, byt ) # rRMSE of test set
    rrmse_train = ComputeError(clf, bx, by ) # rRMSE of tran set
    tr = clf.predict(bx) # training forward pass
    p = clf.predict(bxt) # test forward pass
    v = clf.predict(bxv) # validation forward pass
    ViewPerformance(by, tr, byt, p, byv, v) # plot image
```

Listing 1. Algorithm for generation of enhanced meta models. Optimisation of free parameters with PSO and selection of the best model with minimum error on the validation set.

3.2 Real-Time Design Assessment in Revit

Having established enhanced meta models, we can use them for the fast prediction of the effect of design actions in Revit. In order to do that, a Dynamo node was developed to calculate the forward pass of the meta model based on the model parameters set in Dynamo. The trained meta model is imported as a text file, while parameters are set using value boxes or sliders as shown in Fig. 4(a). The "meta model prediction" node calculates the forward pass, i.e. the analysis response for the selected design parameters. The results are then displayed in Revit with a "visualisation node". This node is visualising simulation results, highlighting both the contour-fill coloured surface for the output quantities as well as the deformation for the case of surface settlements (which is made more visible using a deformation scale factor).

This simple and intuitive representation is crucial for improving the understanding of tunnelling-induced effects by design engineers who can't afford computationally expensive numerical simulations or non-experts who may be involved in the decision process of the project development. Apart from surface settlements, the risk on existing infrastructure is also visualised as shown in Fig. 4(b), where the buildings are sorted in relative scale from green (safe) to red (in risk) based on the influence of the tunnel excavation [21].

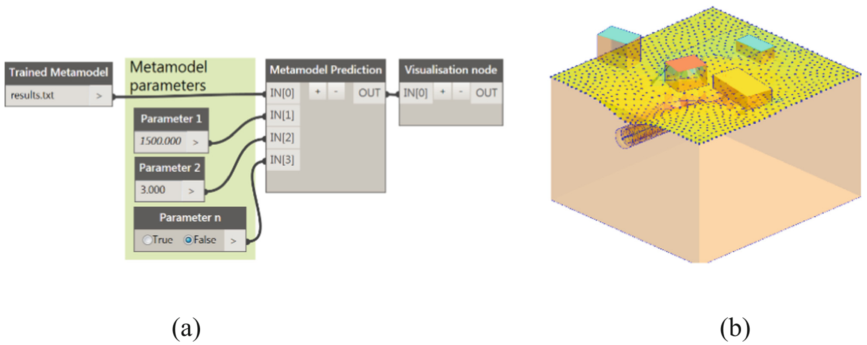


Fig. 4. (a) Dynamo node for real-time design assessment based on meta models; (b) visualisation of numerical results in Revit using contour-fill coloured and deformed surface for surface settlements and contour-fill coloured surface for risk of tunnelling on buildings [18]. (Color figure online)

3.3 Sensitivity Analysis Based on Meta Models

Calculation of the sensitivity measures is implemented in Python as shown in Listing 2. For the forward calculation of the effect of parameter variation onto the output, the meta models trained by means of finite element simulations are applied. This offers significant advantages, since the calculation of the elementary effect requires a large number of forward calculations to ensure the uniqueness of the solution. As explained in the previous section, meta models are excellent tools for interpolation and extrapolation of the trained data set, and therefore a reasonable solution for the forward calculations in this manner. In Listing 2, the following steps are performed:

- define the set of parameter ranges to be investigated (*bxt*);
- use the meta model to predict the model output (*sens_test*) based on given input (*bxt*) and trained meta model synaptic weights (*weights*) using the *ForwardPass()* method;
- calculate the μ_j^* (mean of the absolute value of EE_i) using the *ComputeVariance()* method;
- calculate the σ (standard deviation of EE_i) using the *ComputeDeviation()* method;

```
"Calculate sensitivity measures for data set "bxt"."
def main():
"Read meta model."
   w = ReadWeights(weights)
   sens = []
   devs = []
   bxt = [[0 for x in xrange( input_size)] for x in xrange(n_delta*time_steps)]
      for i in range (0, n_delta):
         for j in range (0, time_steps):
            bxt[i*time_steps+j][1] = 0.1+0.8/(n_delta-1)*i
            bxt[i*time_steps+j][3] = 0.1+0.8/(time_steps-1)*j

      sens_test = ForwardPass(w, bxt, arch_ann, 1, 'relu')
      variance = ComputeVariance(sens_test, n_delta, time_steps)
      dev = ComputeDeviation(sens_test, n_delta, time_steps)
```

Listing 2. Algorithm for sensitivity analysts based on meta models.

4 Examples

4.1 Real-Time Prediction

In the example presented in this paper, we consider the problem of a tunnel passing in the vicinity of an existing building (see Fig. 5). The investigated design and modelling parameters are: (i) building LoDs (1, 2 and 3); (ii) the distance of building's centreline from the tunnel in Y direction (0, 10, 20, 30, 40, 50, 60 m) and; (iii) the tunnel overburden (5, 10, 15, 20, 25 m). SatBimModeller is used for the generation and execution of a large number of simulations in order to obtain the data set for meta model training. In this numerical experiment containing 105 simulations for the construction of 25 tunnel rings with combination of three parameters (building LoD, overburden and distance), the selected monitoring quantities are the vertical displacements of the building top. This results in a data set of 2554 monitoring samples. This set is used for the generation of the meta model for prediction of building settlement w.r.t building LoD, overburden and distance. In the next step, this meta model is used for the forward calculations of sensitivity analyses.

Meta Model Training for the Prediction of Building Settlement. The procedure and the algorithm described in Sects. 2.2 and 3.1 are applied for the meta model training based on a data set of 2554 samples. Here, the data set is divided into portions of 80%, 15% and 5% samples for training, testing and validation, respectively.

Figure 6(a) shows the relative Root Main Square Error (rRMSE) for different machine learning techniques applied for the data training. From the figure, it is clear that the optimized ANN shows the best performance for data set training. Figure 6(b)

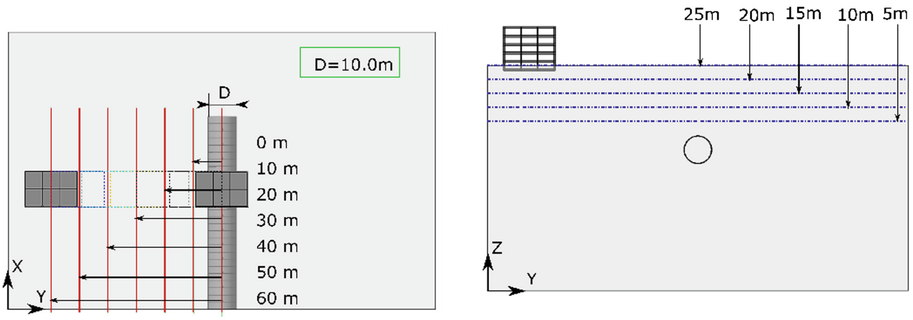


Fig. 5. Parameters for investigation of the building LoD sensitivity. Left: design alternatives in terms of building distance from the tunnel alignment; Right: depth of tunnel w.r.t. foundation of the existing building.

shows the convergence of the optimization of the ANN architecture leading to a minimized error for the test sample. PSO is used as optimization algorithm to determine the number of neurons in the hidden layers and the learning rate. From Fig. 6(b), it can be seen that the optimal solution is reached within approx. 22 iterations leading from non-optimized model with 18% rRMSE to the optimized model with less than 3% rRMSE of the test set.

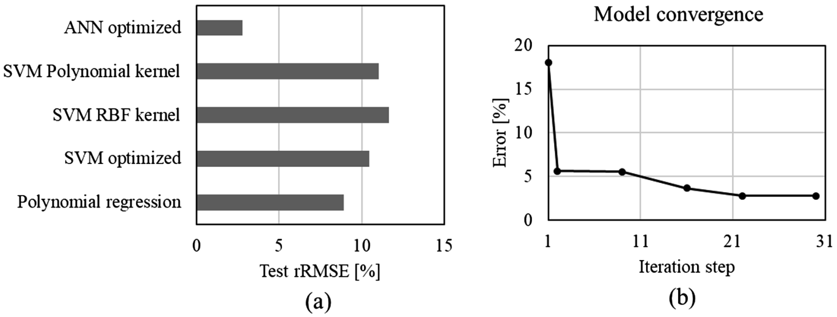


Fig. 6. (a) Training performance of the meta model using the various machine learning techniques; (b) convergence of the optimization process of the ANN architecture.

Figure 7 shows the comparison between the data set and meta model prediction for the training, testing and validation set for the best meta model. From this figure, it can be concluded that the optimized ANN meta model has excellent prediction capabilities, with the error on test and validation set being less than 3%.

On-Demand Design Assessment in Revit. Figure 8 illustrates how simulation-based meta models can be used for real-time prediction and design assessment in Revit. In this example, to assess the impact of tunnel construction on an existing building, the influence of the distance of the tunnel from the existing building and the tunnel depth is

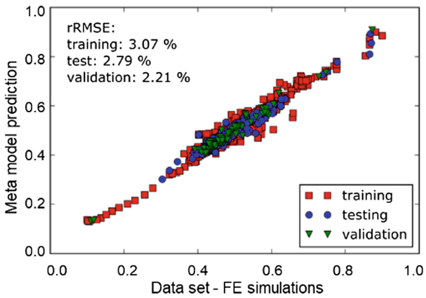


Fig. 7. Comparison between vertical displacements of the building obtained from the FE simulation and predictions of the trained meta model for the training, test and validation sets for the model with best performance

evaluated directly in the design tool. The design parameters are set by user using value boxes and sliders in Dynamo as illustrated in Fig. 8. This design assessment approach can be applied in the early design phase when exploring different design alternatives to minimize the impact of tunnel construction on the existing building. Using the meta model, the results of the analysis are obtained and visualised instantaneously, while the full finite element simulations would have taken hours to calculate. Another advantage of using meta models for real-time design assessment is that meta models are able to interpolate and to certain extent extrapolate the prediction for the explored range of parameters. Thus, for a discrete number of test simulations characterised with a given range of input parameters, using the meta models, the response can be obtained for an infinite number of (continuous) parameter combinations within this range.

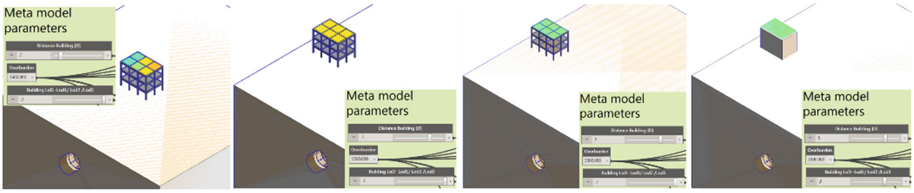


Fig. 8. On-demand design assessment in Revit based on simulation-trained meta models for different tunnel offset and tunnel depth from the existing building and LoD of building model.

4.2 Evaluation of the LoD Importance

A variance-based global sensitivity analysis has been conducted in order to measure the sensitivities of the model output (settlements at the building top) to the input parameters (building LoD, overburden and distance). In this methodology, the importance of the input parameter is quantified through two sensitivity measures σ and μ^* as explained in Sect. 2.3.

In Figs. 9 and 10, the sensitivity measures for the selected LoD of the building to vertical displacements are shown. From the plots, we can conclude that the global

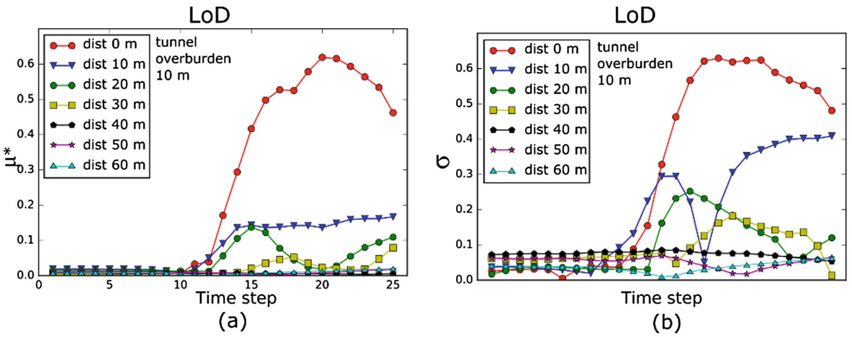


Fig. 9. Sensitivity of the building LoD for different building distance from the tunnel alignment for tunnel overburden of 10 m: (a) absolute mean of EE_i; (b) standard deviation of EE_i.

sensitivity μ^* as well as the interaction effect indicated by σ of the LoD drops with the increase of the distance of the building from the tunnel and the increase of the overburden of the tunnel. It is for instance obvious that when the distance of the building from the tunnel is approximately 4D, the selected building LoD becomes irrelevant, meaning that we can choose the lowest LoD and reduce computational costs.

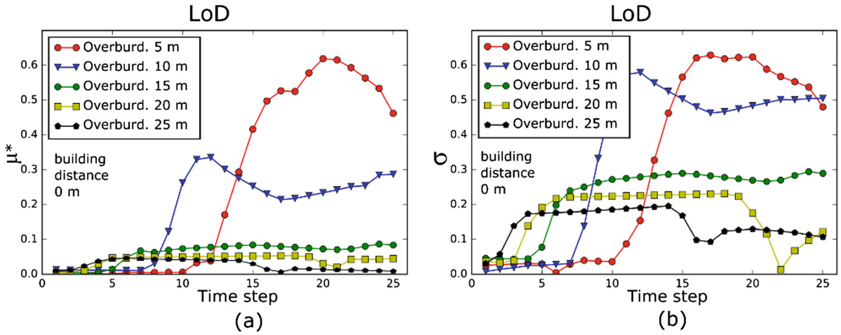


Fig. 10. Sensitivity of the building LoD for different tunnel overburden for building distance of 0 m: (a) absolute mean of EE_i; (b) standard deviation of EE_i.

Looking more closely at the results of the sensitivity measures σ and μ^* of the building LoD for different values of tunnel overburden (Fig. 10), we can see that both the global sensitivity and the interaction effect reduce with the increase of the overburden. However, these sensitivity measures are still significant even for the overburden of 25 m - especially σ, which detects a nonlinear relation between the input and the output. This is due to two reasons: first, because the chosen limit of the overburden of 25 m (2.5 D) from the tunnel crown is still in a zone of large influence, and second because of the building distance from the tunnel in Y direction of 0 m, where the interaction effect is the strongest (see Fig. 9).

5 Conclusions

In this paper, a concept for on-demand tunnelling design assessment in an engineering design environment is proposed. To this end, simulation-based meta models, trained a priori with complex 3D numerical simulation models, are employed for real-time prediction. The unified design-analysis- assessment platform SATBIM is used as a design tool, and as a basis for generation of a large number of simulations for creating a data set for meta model training. Moreover, a strategy and algorithm for generation of enhanced meta models based on different machine learning techniques is proposed. In the example given in this paper, we demonstrated on-demand design assessment of effects of tunnelling on existing building in Revit. Finally, meta models are applied for sensitivity analysis to explore the importance of model parameters.

In general, real-time predictions are required if a large number of alternatives has to be explored and if decisions have to be made in real-time (e.g. setting support pressures during the tunnel construction process). Meta models have been chosen in this approach to substitute complex 3D numerical simulation, since they have been recognised by the practitioners as an efficient method which compromises complexity and speed of calculation [22]. Meta models are able to account for the individual behaviour of each component and their complex interactions, giving a more physical response. Consequently, different design assessment measures can be evaluated at the same time (surface settlements, risk on buildings, stresses in tunnel structure). Certainly, to generate these meta models, a large number of simulation runs has to be performed a priori, and these calculations will require a significant amount of time. However, the advantage is that the generation of simulations is automatized and instantaneous by applying SatbimModeller and that the used simulation models can be parallelised [23]. Secondly, since meta models are able to interpolate and to a certain extent extrapolate the prediction from given parameter ranges, one can test an infinite number of parameter combinations within the chosen range from a discrete number of simulations used for meta model training. Another tradeoff is that trained meta models are characterised with a certain prediction error. Therefore, different machine learning methods and optimization were used here in order to ensure the best training performance. Consequently, in the example given in this paper, the prediction error is less than 3% and therefore acceptable for most engineering applications.

Finally, another important application of meta models, the sensitivity analysis and evaluation of the model output, is demonstrated by an example. Here, meta models are proved useful for fast prediction, since a large number of forward calculation have to be performed to obtain sensitivity measures. By applying meta model-based sensitivity analysis, we evaluated the importance of the building model LoD for numerical assessment. The results can be used in the future to select optimal LoD of building component. This then would lead to optimal information and numerical models in terms of model size and computational efforts.

The SATBIM toolkit will be made available as open source software together with tutorials, a complete manual, and a number of benchmark examples. The project's Github repository (not yet public) can be found at: https://github.com/satbim.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 702874 "SATBIM—Simulations for multi-level Analysis of interactions in Tunnelling based on the Building Information Modelling technology". This support is gratefully acknowledged.

References

1. Potts, D., Zdravkovic, L.: Finite Element Analysis in Geotechnical Engineering: Application. Thomas Telford Ltd., London (2001)
2. Ninić, J., Meschke, G.: Model update and real-time steering of tunnel boring machines using simulation-based meta models. Tunn. Undergr. Space Technol. **45**, 138–152 (2015)
3. Khaledi, K., Miro, S., König, M., Schanz, T.: Robust and reliable metamodels for mechanized tunnel simulations. Comput. Geotech. **61**, 1–12 (2014)
4. Legatiuk, D., Theiler, M., Dragos, K., Smarsly, K.: A categorical approach towards metamodeling cyber-physical systems. In: Proceedings of the 11th International Workshop on Structural Health Monitoring (IWSHM). Stanford, 9 December 2017
5. Smarsly, K., Theiler, M., Dragos, K.: IFC-based modeling of cyber-physical systems in civil engineering. In: Proceedings of the 24th International Workshop on Intelligent Computing in Engineering (EG-ICE). Nottingham, 7 October 2017
6. Schulz1, W., Hermannsa, T., Khawlia, T.A.: Meta-modelling, visualization and emulation of multi-dimensional data for virtual production intelligence. AIP Conf. Proc. **1863** (2017). 440003
7. Suwansawat, S., Einstein, H.: Artificial neural networks for predicting the maximum surface settlement caused by EPB shield tunneling. Tunn. Undergr. Space Technol. **21**(2), 133–150 (2006)
8. Kim, C.Y., Bae, G., Hong, S., Park, C., Moon, H., Shin, H.: Neural network based prediction of ground surface settlements due to tunnelling. Comput. Geotech. **28**(6–7), 517–547 (2001)
9. Borrmann, A., Flurl, M., Jubierre, J.R., Mundani, R.-P., Rank, E.: Synchronous collaborative tunnel design based on consistency-preserving multi-scale models. Adv. Eng. Inform. **28**(4), 499–517 (2014)
10. Koch, C., Vonthron, A., König, M.: A tunnel information modelling framework to support management, simulations and visualisations in mechanised tunnelling projects. Autom. Constr. **83**, 78–90 (2017)
11. Meschke, G., Freitag, S., Alsahly, A., Ninic, J., Schindler, S., Koch, C.: Numerical simulation in mechanized tunneling in urban environments in the framework of a tunnel information model. Bauingenieur **89**(11), 457–466 (2014)
12. Ninić, J., Koch, C., Stascheit, J.: An integrated platform for design and numerical analysis of shield tunnelling processes on different levels of detail. Adv. Eng. Softw. **112**, 165–179 (2017)
13. Ninić, J., Koch, C., Tizani, W.: Parametric information modelling of mechanised tunnelling projects for multi-level decision support. In: 24th EG-ICE International Workshop on Computing in Engineering, vol. 1, pp. 228–238 (2017)
14. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis, 5th edn. Wiley, New York (2012)
15. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
16. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (2000). https://doi.org/10.1007/978-1-4757-3264-1

17. Kennedy, J., Eberhart, R. C.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, Piscataway, pp. 1942–1948 (1995)
18. Morris, M.: Factorial sampling plans for preliminary computational experiments. Technometrics **33**, 161–174 (1991)
19. Miro, S., Hartmann, D., Schanz, T.: Global sensitivity analysis for subsoil parameter estimation in mechanized tunneling. Comput. Geotech. **56**, 80–88 (2014)
20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikitlearn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
21. Schindler, S., Hegemann, F., Koch, Ch., König, M., Mark, P.: Radar interferometry based settlement monitoring in tunnelling: Visualisation and accuracy analyses. Vis. Eng. **4**, 7–23 (2016)
22. Stascheit, J., Ninić, J., Hegemann, F., Maidl, U., Meschke, G.: Building information modelling in mechanised shield tunnelling – a practitioner's outlook to the near future. Geomech. Tunn. **11**(1), 34–49 (2018)
23. Bui, H.G., Alsahly, A., Ninic, J., Meschke, G.: BIM-based model generation and high performance simulation of soil-structure interaction in mechanized tunnelling. In: The Fifth International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering (PARENG 2017), Paper 44. Civil-Comp Press, Stirlingshire (2017). https://doi.org/10.4203/ccp.111.44

A Visual Interactive Environment for Engineering Knowledge Modelling

Ewa Grabska, Barbara Strug, and Grażyna Ślusarczyk(✉)

Jagiellonian University, Kraków, Poland
{ewa.grabska,grazyna.slusarczyk}@uj.edu.pl

Abstract. Today, the field of modelling of engineering knowledge needs facilitating the specification and programming tasks associated with the modelling of complex systems. In this paper a graphical formalism is suggested as a way to tackle the system modelling problems more quickly. Composition graph (CP-graph) rewriting is suggested as a basis for visual modelling tools. Nodes of a CP-graph have explicit connection elements called bonds to which edges are attached. The proposed Visual MOdelling with GRaphs (VIZMOGR) system offers a simple graphical interface for conceptual designing artifacts, in which outline solutions are determined by spatial and structural relationship of the principal components and major functions. The system allows the user to create structure components of the model in the form of CP-graph nodes labelled by icons representing artifact's components. Creating edges between bonds of nodes is controlled by means of label-dependent rules. To modify created CP-graphs VIZMOGR system provides schemes of rules, which are most often applied to develop graphs. Moreover VIZMOGR system supports attributed CP-graphs by enabling the user to propagate semantic information and to capture parametric modeling knowledge. The benefits of the proposed approach and usefulness of VIZMOGR system are shown using examples of designing bridges.

Keywords: Knowledge modelling · Composition graph · Graph rewriting

1 Introduction

Today, the modelling of knowledge is made very complex by the use of multiple discipline-specific models. Integration of these models is difficult because different scientific bases are applied [1]. We are at a point in history when there exists an obvious need for facilitating the specification and programming tasks associated with the modelling of complex systems.

We consider here the process of modelling corresponding to early stages of design, called conceptual design, where primary decisions are to be made. The product of these stages is meant as an outline solution to a design problem in which spatial and structural relationships of the principal components are fixed, and major functions are determined. The outline solution can be easily represented by a graph [2].

In this paper a graph formalism is suggested as a way to tackle the system modelling problems more quickly. The consequence of using this formalism is stimulation of visual

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 219–230, 2018.
https://doi.org/10.1007/978-3-319-91635-4_12

thinking that means taking advantage of our innate ability to see, discovering ideas that are otherwise invisible, developing these ideas quickly and intuitively, and then sharing them with others. *Composition graph (CP-graph)* rewriting is proposed as a basis for visual modelling tools. Intuitively, a CP-graph is a graph where nodes have explicit connection elements called *bonds* [3]. Edges are attached to bonds which express the properties of the connections (see Fig. 1). Graph rewriting consists of replacing a subgraph in the graph being processed by another graph, thus giving the possibility of creating a new graph out of the original graph algorithmically. We focus on CP-graph rewriting systems that have been used to synthesize models in a wide variety of domains such as architectural design, computer games, computational grids and Finite Element Method [4, 5].

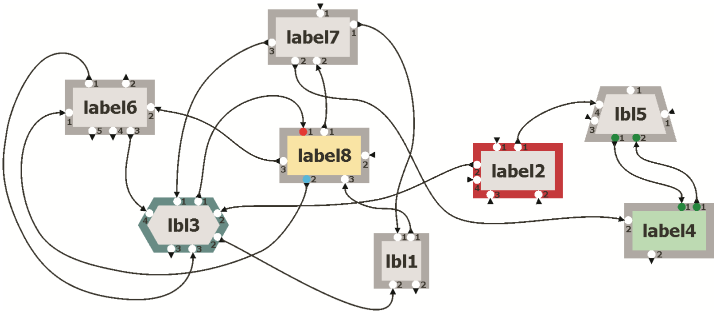


Fig. 1. An example of a CP-graph structure.

There exist a number of tools for generating graphs such as PROGRESS, AGG, Fujaba, GROOVE, but they are usually specialized for experts of a particular subject and focus on different areas. An interesting tool for graph rewriting is PORGY, which uses port graphs [6]. Ports, like bonds, are connection elements between edges and nodes defined on the syntactic level. However, the difference between ports and bonds is fundamental, since bonds are also interpreted on the semantic level. This interpretation plays an essential role in building a layer of meaning during the initial phase of modelling, called conceptualization.

The existing systems can be considered in the framework of Computational Design Synthesis (CDS). CDS aims to support conceptual design through formalization and computer aided process of finding solutions for knowledge-based design tasks. The lack of popularity of CDS in industry is evident. This lack of acceptance is caused by a great effort required to acquire knowledge and skills necessary to formalize tasks, limited range of applications, lack of reuse of existing design knowledge, lack of modeling standards and tool integration [7]. For instance, CDC is often discussed within the framework of local graph transformations, specified by graph rewriting rules for transforming subgraphs of limited sizes. The graph approach is fundamental because it allows to consider both the product modelling and the process modelling. There exist many formal graph approaches to the application of graph rewriting but their use most often requires skills to formalize tasks. The main purpose of this paper is to facilitate the specification of graph rewriting rules by presenting steps of applying them on the visual level.

The proposed VIZMOGR (Visual MOdelling with GRaphs) system is a visual environment that allows users to define CP-graphs and rewriting rules, and to apply these rules in an interactive way or via the use of a strategy with memory, where successive rewriting steps depend on the history of the previous ones [8]. VIZMOGR system offers a simple graphical interface that allows users to visualize and analyse the dynamics of finding solutions for knowledge-based design tasks. The graph rewriting in the system is based on the GraphTool engine [4, 5]. In the first step of modelling, when the user analyses relevant entities and organizes them into concepts and relations, he can decide on shapes and colours of elements creating a graph structure of the model. It turns out that visual perception of graph structures in graphic design is based on the same neural machinery that is used to interpret the everyday environment [9]. A kind of natural semantics, which is fundamental in reasoning, is built on the basis of user's patterns.

VIZMOGR system offers the support for attributed CP-graphs. Within each project there is a file that is used for creating declaration of attributes that can be later assigned to any object defined in the project. Each rewriting rule can be equipped with a predicate of applicability specifying conditions under which the rule can be used. These predicates are in the form of expressions with parameters being attributes assigned to CP-graph nodes and edges and therefore describe semantic properties of CP-graphs to which rules can be applied.

The benefits of the proposed approach and usefulness of VIZMOGR system are shown using examples of designing bridges.

2 CP-Graph Representation of Designs

In the proposed VIZMOGR system design objects are internally represented by means of CP-graphs. A CP-graph is a labelled and attributed graph, where nodes represent components of artefacts and are labelled by names of components they represent. Moreover, attributes specifying properties of components are assigned to nodes representing them. To each node a number of bonds expressing potential connections between components is assigned. Edges of a CP-graph represent relations between components. Non-symmetrical relations are represented in a directed CP-graph, where nodes are equipped with two types of bonds: in-bonds and out-bonds and directed edges are drawn from out-bonds to in-bonds. Figure 1 shows an example of a generic CP-graph structure with directed edges, which represents a designed model. In this CP-graph, for instance the node "*label7*" and the node "*lbl3*" are connected by the directed edge which is drawn from the out-bond of the former node to the in-bond of the latter one.

Bonds of CP-graph nodes, which are not engaged are called free and represent places on components where other components can be attached. In Fig. 1 the node *"label7"* has one free in-bond, whereas the node "*lbl3*" has one free out-bond. Each free bond of a node signals a potential connection of a node with other nodes [3]. Undirected CP-graphs have only one type of bonds. In Fig. 2b a CP-graph representing the structure of the bridge shown in Fig. 2a is undirected, i.e., its edges represent symmetrical relations. This structure is a solution internal representation obtained as a result of the modeling process controlled by the user and described in the next section. Nodes of this CP-graph represent abutments, pylons, beams and two types of cables, which constitute structural components of the bridge, and are labeled as *ab*, *bm*, *pl*, *cb1* and *cb2*, respectively. Bonds assigned to nodes are connected by undirected edges representing relations between bridge components. The edge labels *fx* and *hn* denote fixed and hinged connections between components of the bridge. The edge label *jn* denotes the join relation.

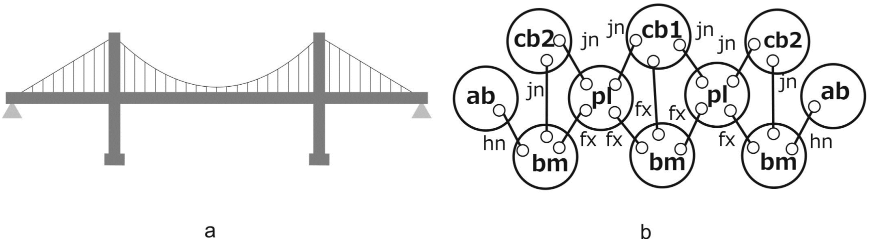


Fig. 2. A bridge and a CP-graph representing it.

3 VIZMOGR System Characteristics

The proposed system offers a simple graphical interface that allows users to visualize and analyse the dynamics of solutions for knowledge-based design tasks. The system enables the user to create CP-graphs corresponding to design drawings being early solutions.

At first, in the conceptualization phase, the user specifies icons representing object structural components the design drawings are to be generated of together with their labels. In this way a set Σ_V of labels which correspond to the icons is determined. Then, a set of graph nodes labelled by icons is created.

Exemplary icons, which are used in case of designing bridges to represent structural components of models, together with their labels are presented in Fig. 3a. Therefore set Σ_V contains labels *ab*, *pl, bm*, *sp*, *ar1, ar2, cb1, cb2* and *cb3*, which represent abutments, pylons, beams, supports, two types of arches, and three types of cables.

When we see CP-graph nodes with icons we are relying on the same neural machinery that is used to interpret the everyday environment. This natural semantics permeates the visual thinking. Abstract phrases such as *connected* or *contained within* are not considered metaphoric. In our approach the interplay between CP-graph nodes with icons is considered. Two CP-graph nodes connected by edges represent related components. The places of connections are represented by small circles in nodes called *bonds*. CP-graph nodes can contain bonds which are not connected by edges. They are called free and signal potential connections of a node with other nodes. CP-graph nodes with labels in the form of icons and with bonds assigned to them are shown in Fig. 3b.

Visual relations which can occur between icons are also specified and a set Σ_E of labels representing types of possible relations between object components, is determined. For bridges, the set Σ_E contains labels *fx, hn* and *jn*, which represent relations corresponding to fixed and hinged connections between components of bridges, and the join relation, respectively. When the user analyses relevant entities and organizes them

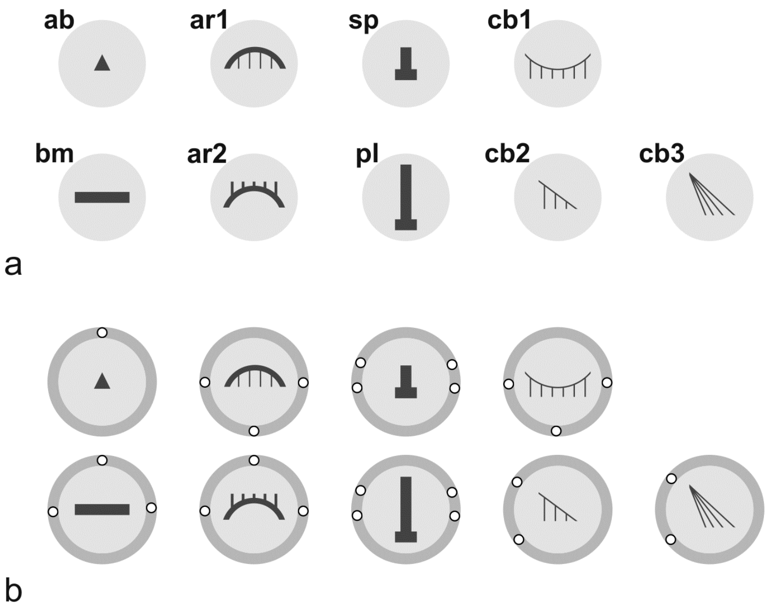


Fig. 3. (a) Icons representing components of bridges, (b) CP-graph nodes representing these components.

into concepts and relations, he can decide on shapes and colours of elements representing graph nodes. Moreover, for each node a set of attributes specifying properties of a corresponding component can be specified.

The CP-graph is created by adding selected nodes and connecting their bonds by edges representing relations between components corresponding to nodes. It is worth noting that starting with designing a fragment of the bridge structure in the form of a configuration of nodes with icons, the location of individual nodes can be easily determined by the relations (for instance *below*, *on the left*) among graphical primitives in the bridge drawing (see Fig. 4).

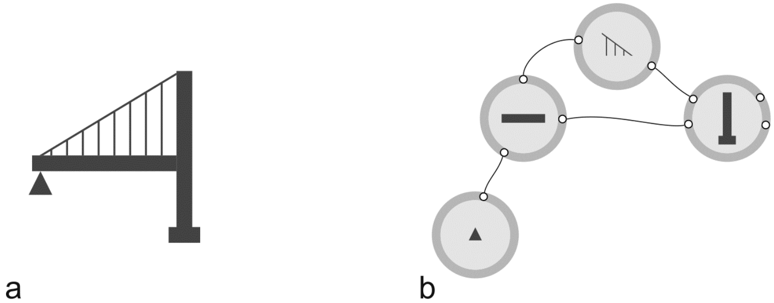


Fig. 4. (a) A fragment of a bridge (b) a CP-graph corresponding this fragment.

Creating edges between bonds of nodes is controlled by means of label-dependent rules. If the created edge is directed, its source and target bonds are automatically converted into out-bonds and in-bonds, respectively. The system gives also the possibility to label edges, remove nodes and edges, and change labels of both nodes and edges.

In Fig. 5 the first screen of GUI of VIZMOGR system is presented. In the left-hand side window the available icons with their labels are shown. In the right-hand side window a CP-graph representing a preliminary design can be created.

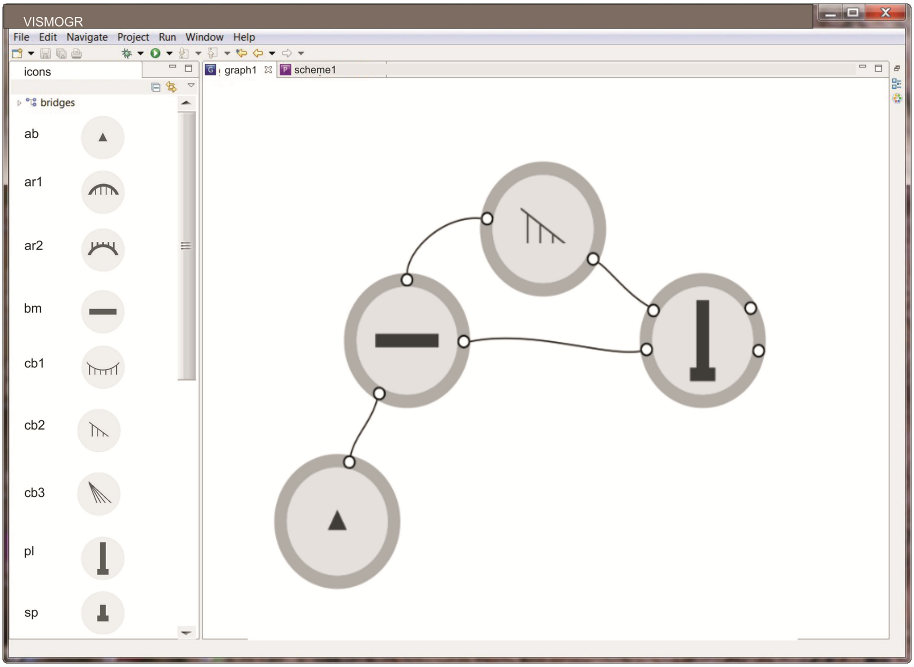


Fig. 5. A screenshot of the first part of VIZMOGR GUI, where CP-graphs can be created.

CP-graphs representing design drawings can be further modified by automatically applying sequences of CP-graph transformation rules selected by the user. The graph rewriting is based on the GraphTool engine [4, 5].

A CP-graph rewrite rule consists of two CP-graphs L and R (see Fig. 6a). The former (L) describes a subgraph of a derived CP-graph that after application of the rewriting operation is replaced by the latter (R), which is embedded in the rest CP-graph. Bonds of CP-graph nodes, which are not connected with edges are called free. They play an important role in the rewriting operation. Graphs L and R with the same number of free bonds can easily replace each other with the constant embedding in the rest part of the CP-graph.

Figure 6b shows a visualization of the rewriting rule with constant embedding for L and R. The rhombus, in which to each red solid edge one blue dotted edge is assigned, indicates a bond of R which replaces a bond of L in the derived CP-graph. VIZMOGR system allows users to define rewriting rules with other embeddings by drawing red

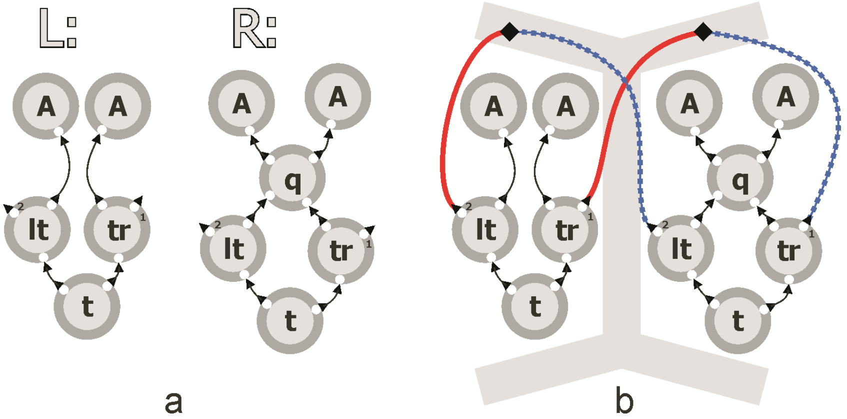


Fig. 6. (a) Two CP-graphs *L* and *R*, (b) a visualization of embedding. (Color figure online)

solid edges with rhombuses for all free bonds of the CP-graph *L* and drawing blue dotted edges connecting free bonds of the CP-graph *R* with appropriate rhombuses.

In order to modify created CP-graphs, firstly CP-graph transformation rules are specified by the user. VIZMOGR system offers five schemes of rules, which are presented in Fig. 7. These schemes correspond to rules which are most often applied in order to develop graphs. The rule presented in Fig. 7a allows for adding a new node and connecting its bond by an edge with a bond of the existing node. The rule from Fig. 7b allows for adding a new node and connecting its bonds by edges with bonds of the two already existing nodes. The rules shown in Figs. 7c–e allow for adding two new nodes with bonds connected by an edge and connecting bonds of them with bonds of the existing node. The difference in rules from Fig. 7c–e lies in location of free bonds in the nodes of the rule right-hand sides. The embedding of the scheme rules is indicated using rhombuses. The user has also the possibility to define his own rule schemes.

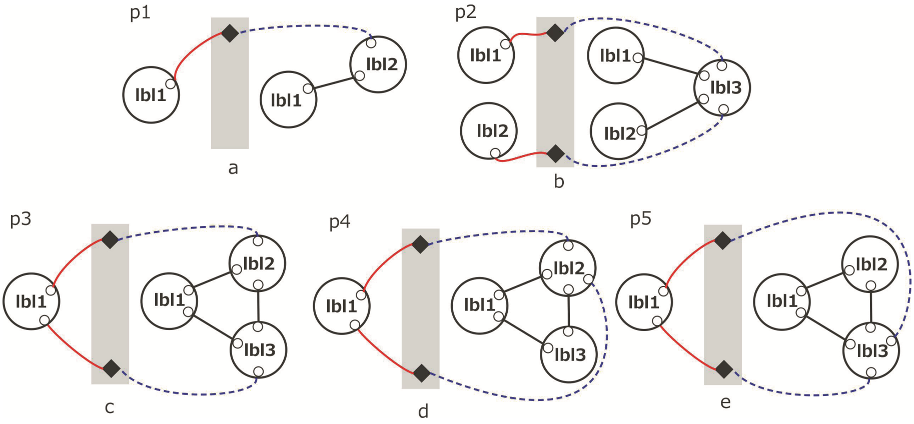


Fig. 7. Five CP-graph transformation rule schemes.

After selecting a given rule scheme the user gives labels of Σ_V to nodes and labels of Σ_E to edges, specifies numbers of bonds for nodes, and direction of edges. Schemes adapted for modelling bridges are presented in Fig. 8. The first and third scheme (*p1* and *p3*) have been adapted twice, while schemes *p2* and *p5* only once. It should be noted that after schemas adaptation the right and left-hand sides of rules can have different number of free bonds. This is admissible, as the embedding transformation is defined by rhombuses. We assume that the nodes of the left-hand side of a rule are found in a derived CP-graph if they have at least the same number of free bonds in this CP-graph. The other bonds, either engaged or free, assigned to the matched nodes remain unchanged during the rule application.

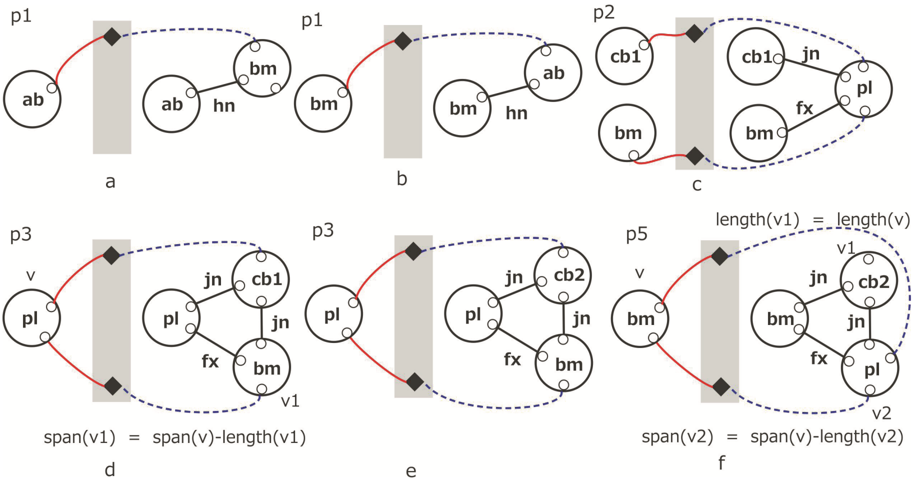


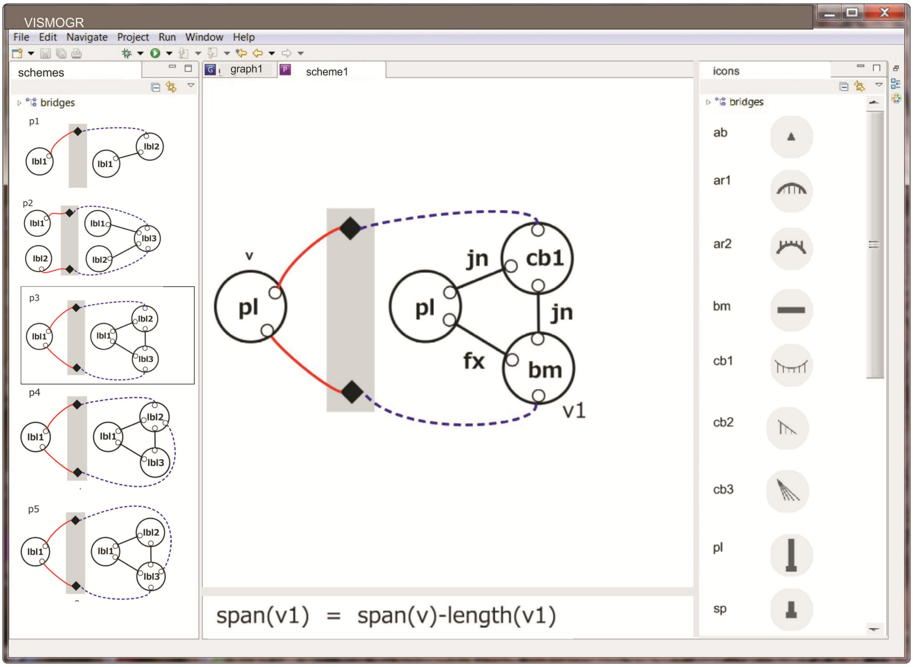
Fig. 8. CP-graph transformation rule schemes adapted for bridge modelling.

Applying CP-graph transformation rules leads to generation of different CP-graphs representing possible design object structures with geometry and material properties specified by graph attributes. There often exists a need to preserve or even to propagate some information throughout the modelling process. VIZMOGR system offers the support for attributed CP-graphs. Within each project attributes can be declared and assigned to CP-graph elements. The user defines the way of transferring the values of attributes from the left to the right hand side. The possibility of relating attributes of right-hand sides of CP-graph rules to attributes of their left-hand sides enables us to propagate semantic information and also to capture parametric modeling knowledge.

In case of bridge design to nodes of both sides of rules the attributes *length* and *span,* which specify the length of the corresponding bridge fragments and the remaining span of the bridge, respectively, are assigned. The initial value of the *span* attribute is decreased after adding nodes representing new bridge elements. For example in the rule presented in Fig. 8d the value of the attribute *span* of the node *v1* labelled *bm* on the rule right-hand side is equal to the value of the attribute *span* assign to the node *v* of the left-hand side decreased by the length of the beam represented by the node *v1* of the right hand side. In the rule presented in Fig. 8f the value of the attribute *length* of the node *v1* labelled *cb2* on the rule right-hand side has to be the same as the value of the same attribute assigned to node *v* of the rule left-hand side.

Each rewriting rule can be equipped with a predicate of applicability specifying conditions under which the rule can be used. These predicates are in the form of expressions with parameters being attributes assigned to CP-graph nodes and edges and therefore describe semantic properties of CP-graphs to which rules can be applied. For example the rule, which corresponds to elongating the bridge by adding a new fragment (see Fig. 8d), can be applied under the condition that the remaining span of the bridge is longer than 100 m, i.e., *span(v) > 100*. Moreover, the rewriting rules can access and/or modify so called memory which is a common set of variables. Then successive rewriting steps can depend on the history of the previous ones.

In Fig. 9 the screen of GUI of VIZMOGR system, where CP-graph transformation rule schemes can be adapted to a given application domain, is presented. In the left-hand side window the available rule schemes are shown. The selected one is marked by the black frame. The same scheme during the adaptation process is shown in the central window. In the right-hand side window icons with their labels, which can be used in the adaptation, are presented. In the bottom window the values of attributes assigned to CP-graph nodes can be specified.



span(v1) = span(v)-length(v1)

Fig. 9. A screenshot of the second part of VIZMOGR GUI, where CP-graph transformation rules can be adapted.

CP-graph transformation rules applied to initial CP-graph representations of drawings generate structures representing new designs. In VIZMOGR system the user can easily define the strategy of controlling CP-graph rewriting process, by drawing a graph, called control diagram, which specifies the possible order of applying CP-graph rules. The ordered sequences of rules, which can be obtained in this diagram, reflect the course of the design process.

In Fig. 10 a control diagram, which specifies the possible order of applying CP-graph transformation rules shown in Fig. 8, is presented. Rule numbers *p1* to *p6* correspond to rules illustrated in Figs. 8a to f, respectively. Starting from the node labelled *ab* and applying rules in the sequences *p1,p2*; *p1,p6,p5,p2*; *p1,p6,p4,p3,p5,p2* and *p1,p6,p4,p3,p4,p3,p4,p3,p5,p2* results in CP-graphs representing bridges from Figs. 11a–d, respectively.

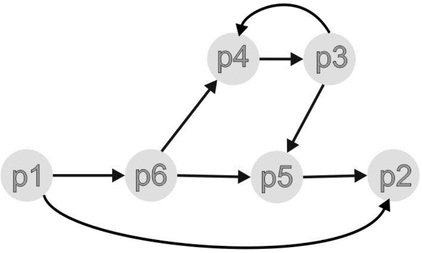


Fig. 10. A control diagram for bridge design.

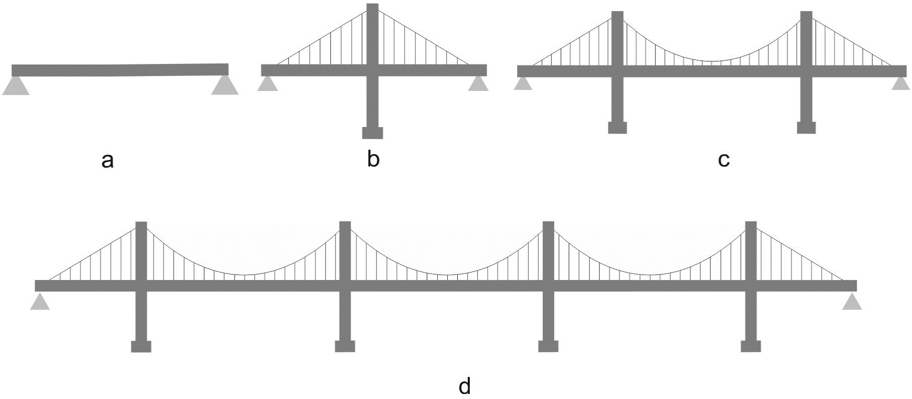


Fig. 11. Bridges represented by CP-graphs obtained using different sequences of rules.

The proposed VIZMOGR system is easy to adapt to a wide range of applications, as CP-graph elements and rule schemes are defined in a generic way. Sets of labels, nodes with bonds and transformation rules can be defined for multiple applications in different domains. The system is easy to use as the structure of available schemes of CP-graph transformation rules can be intuitively understood. Moreover, new schemes can be added in a simple way. Therefore, the proposed approach efficiently supports the task of encoding various types of knowledge. The system also facilitates the specification of CP-graph rewriting rules by presenting steps of applying them in the visual way.

4 Conclusion

In this paper the prototype interactive system which supports modelling of engineering knowledge has been described. In the unified graphical environment the user can select icons representing object structural components and determine relations between them in order to obtain preliminary designs. On the basis of these designs their internal representations in the form of CP-graphs are obtained automatically. The conceptual phase of the object design process is also supported by providing schemes of CP-graph transformation rules in a visual editor. These schemes are adapted for a given application domain and used for modifying graph structures by applying sequences of rules selected by the user. In this way different CP-graphs representing possible design object structures with geometry and material properties specified by graph attributes are obtained. The presented examples show the way in which the system supports encoding knowledge needed to create bridge designs.

In the next steps of our research the catalogues of icons for various design domains will be developed. The memory of graph rules will be used to ensure the presence of a predefined number of selected components within a designed object and to preserve some required characteristics of the design. Moreover the system should enable the user to specify design requirements, and then by checking obtained CP-graph structures verify satisfaction of these requirements by the created design solutions.

References

1. McMahon, C.: Open issues in design informatics. In: Proceedings of the International Conference on Methods & Tools for CAE – Concepts and Applications, pp. 7–12. Bielsko-Biała (2017)
2. Helmes, B.: Object-Oriented Graph Grammars for Computational Design Synthesis (Dissertation), Institution Technische Universität München (2013)
3. Grabska, E.: Graphs and designing. In: Schneider, H.J., Ehrig, H. (eds.) Graph Transformations in Computer Science. LNCS, vol. 776, pp. 188–202. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57787-4_12
4. Palacz, W., Paszyńska, A., Świderska, I., Ślusarczyk, G., Strug, B., Grabska, E.: GraphTool: a visual support for generating graph models of artefacts. In: Proceedings of the International Conference on Methods & Tools for CAE – Concepts and Applications, pp. 81–86. Bielsko-Biała (2017)
5. Palacz, W., Paszyńska, A., Świderska, I., Ślusarczyk, G., Strug, B., Grabska, E.: GraphTool: case studies. In: Proceedings of the International Conference on Methods & Tools for CAE – Concepts and Applications, pp. 75–80. Bielsko-Biała (2017)
6. Fernandez, M., Kirchner, H., Pinaud, B.: Strategic Graph Rewriting: an Interactive Modelling Framework, [Research Report]. Inria, LaBRI, King's College London (2017)

7. French, M.J.: Conceptual Design for Engineers, 3rd edn. Springer, London (1999)
8. Strug, B., Paszyńska, A., Paszyński, M., Grabska, E.: Using a graph grammar system in the finite element method. Int. J. Appl. Math. Comput. Sci. **23**, 839–853 (2013)
9. Ware, C.: Visual Thinking for Design. Morgan Kaufmann, Elsevier (2008)

Lessons Learned with Laser Scanning Point Cloud Management in Hadoop HBase

Anh-Vu Vo[1], Nikita Konda[2], Neel Chauhan[1],
Harith Aljumaily[1,3], and Debra F. Laefer[1(✉)]

[1] New York University, Brooklyn, NY 11201, USA
debra.laefer@nyu.edu
[2] University at Buffalo, Buffalo, NY 11260, USA
[3] Carlos III University of Madrid, Madrid, Spain

Abstract. While big data technologies are growing rapidly and benefit a wide range of science and engineering domains, many barriers remain for the remote sensing community to fully exploit the benefits provided by these powerful and rapidly developing technologies. To overcome existing barriers, this paper presents the in-depth experience gained when adopting a distributed computing framework – Hadoop HBase – for storage, indexing, and integration of large scale, high resolution laser scanning point cloud data. Four data models were conceptualized, implemented, and rigorously investigated to explore the advantageous features of distributed, key-value database systems. In addition, the comparison of the four models facilitated the reassessment of several well-known point cloud management techniques founded in traditional computing environments in the new context of a distributed, key-value database. The four models were derived from two row-key designs and two columns structures, thereby demonstrating various considerations during the development of a data solution for high-resolution, city-scale aerial laser scan for a portion of Dublin, Ireland. This paper presents lessons learned from the data model design and its implementation for spatial data management in a distributed computing framework. The study is a step towards full exploitation of powerful emerging computing assets for dense spatio-temporal data.

Keywords: LiDAR · Point cloud · Big Data · Spatial data management · Hadoop HBase · Distributed database

1 Introduction and Background

Three-dimensional point clouds are increasingly considered as important geospatial resources for a vast range of applications. Point clouds are being collected at an unprecedented rate even at national scales [1]. Yet efforts to harness the usefulness of such datasets are increasingly threatened by the data's expanded size, intensified density, and enhanced complexity. Effective storage, querying, and visualization are essential to successfully address these data challenges. While traditional relational database

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 231–253, 2018.
https://doi.org/10.1007/978-3-319-91635-4_13

management systems (RDBMSs) have been in service for decades, the advent of non-relational alternatives offers an attractive new set of options. Specifically, many non-relational data systems are capable of handling the petabytes of data emerging from the Big Data regime. To begin exploring the capabilities of these powerful computing assets, this paper presents an investigation of HBase – a distributed, non-relational, key-value storage platform within the Hadoop ecosystem for point cloud storage and querying.

To achieve this, the good practices established for point cloud data management in traditional environments are implemented and evaluated in the non-relational database context with four hypothetical data models. Throughout the paper, comparisons against previous RDBMS implementations are highlighted. The main aim is to share the lessons learned from the migration from an RDBMS context to a non-relational alternative with the prospect of building an integrated distributed, spatio-temporal database system for urban data. At the time of writing, the system is capable of providing concurrent access to laser scanning point data in the forms of exact match and three-dimensional (3D) range search. Data compression is supported by HBase's in-built compression mechanisms (e.g. Snappy, LZO, GZIP). The query accuracy of range searches can be set at the point or block level so that users can prioritize either accuracy or querying speed. Additional functionalities such as level-of-detail are not currently supported are planned for the future.

To provide the necessary background for the work presented in the paper, the remainder of this section introduces essential concepts behind Big Data and several technologies for handling Big Data including non-relational databases such as HBase on which this paper is based.

1.1 Big Data Challenges and Hadoop Technologies

According to the in-development ISO standard, ISO/IEC DIS 20546, Big Data are datasets of extensive volume, variety, velocity, and/or variability that require scalable technologies for efficient storage, manipulation, management, and analysis. While the specific traits attributable to the nature of Big Data are still a subject to debate [2], the main technological challenge incurred by Big Data is the profound demand on performant and scalable computing power to handle the data's growth in (1) size, (2) accumulation speed, and (3) complexity. The two common solutions to source the increasingly needed computing power involve a more powerful stand-alone computer (i.e. a supercomputer); or distributing the computation over multiple computers (i.e. a computing cluster). The two approaches are referred to as scale-up and scale-out solutions. The scale-out approach, also known as distributed computing, is often more cost effective and more sustainable when data growth is expected to continue. The hardware configuration (i.e. scale-out or scale-up) must be accompanied by an appropriate programming framework. Dominant amongst existing distributed programming paradigms for scale-out computing clusters are the Message Passing Interface (MPI) and MapReduce. MPI suits tightly-coupled problems that require certain intensive communication between computing nodes to share data and the computational states. In contrast, MapReduce (which falls under the shared-nothing category) is restricted to computations that can be decoupled into independent components that require highly limited exchanges between them.

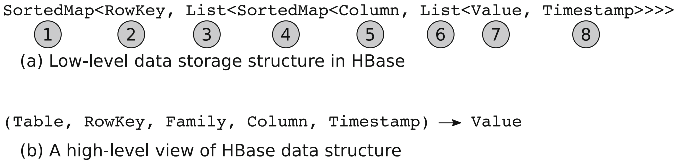
Critical to the recent advancements in Big Data technologies are the increasing popularity of low-cost hardware and open-source software enabling parallel programming. Amongst the existing parallel, distributed computation frameworks, Hadoop is perhaps one of the most familiar names. Hadoop originated from a MapReduce web indexing project lead by Doug Cutting in 2002 that replicated the distributed data storage system and processing framework developed at Google [3, 4]. Today, the name Hadoop is used beyond that initial single project to indicate an entire ecosystem of software and hardware solutions supporting distributed computing on commodity computing clusters. Facebook, Google, Yahoo, IBM are amongst the prominent Hadoop cluster owners, but there is speculation that these powerful computing assets may soon be as accessible as personal computers became in the 1990s. In fact, cloud computing has already made the technology available to anyone with a reliable internet connection and a credit card, irrespective of locale.

Hadoop is neither the only nor the first distributed computing technology. Parallel computing and distributed computation were well developed field long before the emergence of Hadoop. However, Hadoop is amongst the most-used distributed computing technologies today [5]. Other critical distinguishing features behind the popularity of Hadoop is attributable to its accessibility via open-source, its suitability for a wide variety of generic applications, and its friendliness to non-expert users. Hadoop abstracts most of the complexity of distributed computing away from users, while only exposing them to a rather high-level, highly-simplified programming interface. Furthermore, users need not directly handle all of the internal complexity of distributed computation to be able to exploit its power.

1.2 HBase - A Distributed Data Management System Within Hadoop

Within the Hadoop ecosystem is HBase, an open-sourced replica of Google's BigTable [6]. This data management system allows random data retrieval at a petabyte scale distributed over thousands of servers. Unlike the original Hadoop data storage system, which only supports batch processing, HBase allows random access to the distributed data. Since the data are distributed, HBase databases are inherently highly parallelized. Thus, data retrieval is extremely efficient. Compared to traditional relational database management systems (RDBMS), HBase provides much higher flexibility, as it does not require a rigid data schema or even data types. Instead, all HBase data are maintained in their arbitrary binary form and can be interpreted at the time of reading.

At the lower level, HBase data are maintained as a large multi-dimensional sorted map, which can be expressed programmatically as in Fig. 1 [7]. According to that data structure, an HBase table is a sorted map ① of pairs of RowKey ② and List ③. Each element of List ③ is called a column family in HBase. A row key is a user-defined, unique identifier of each row in the HBase table. Notably, the row key plays an important role in HBase indexing as it is the primary means for sorting and also distributing the data. As a result, deciding upon the row key design is of utmost importance in HBase table design, as will be demonstrated in the latter part of this paper. Each column family [i.e. SortedMap ④] is composed of pairs of the table column ⑤ and a list ⑥ of table value and timestamp pairs [i.e. ⑥ and ⑧]. The value is the actual data content stored in the table, while the timestamp denotes the creation time of the content. The timestamp allows storage of multiple versions of the content in HBase. The data structure of an HBase table is sometimes viewed at a higher level as a collection of key-value pairs, in which a key is composed of a row-key, a column family name, a column name, and a timestamp. The value is the actual datum.



```
SortedMap<RowKey, List<SortedMap<Column, List<Value, Timestamp>>>>
```

(a) Low-level data storage structure in HBase

```
(Table, RowKey, Family, Column, Timestamp) ⟶ Value
```

(b) A high-level view of HBase data structure

Fig. 1. HBase's data storage structure

Despite all of its favorable characteristics, HBase is not a replacement for a traditional RDBMS. While aiming for higher performance and greater flexibility, the HBase design (as with most other non-relational database systems) loosens parts of the relational features such as the compliance to Codd's 12 rules and the guarantees against transaction validity (a.k.a. ACID) – the traditional, widely-adopted RDBMS standards [7]. Even though these trade-offs are not acceptable in domains such as banking and medical databases, they are not fatally problematic in many applications such as web searching or point cloud visualization. Another feature that may defer the use of HBase is the lack of capability to model data relations. Notably, each HBase table is independent and contains no explicit relation with other tables. Powerful functionalities in RDBMS including foreign key and join are not inherently supported in HBase. In summary, HBase is introduced in this section as being representative of a new generation of high-performance, highly scalable, cost-effective non-relational data management systems that serve as alternatives to traditional relational databases. While HBase and other non-relational systems surpass traditional RDBMS with respect to many important criteria, they are not the definitive choice in every scenario. The decision between an RDBMS and a more relaxed non-relational option must be based on a rigorous justification of the features of the candidate technologies with respect to the specific data storage and retrieval demands. Some of the rationales for the selection of non-relational solutions for point cloud data storage and management are presented in Sect. 1.3.

1.3 Laser Scanning Point Cloud as a Growing Source of Big Data

One fast growing area where a Big Data solution is clearly needed is in the storage of Light Detection And Ranging (LiDAR) data. The LiDAR technology (also known as laser scanning) [8, 9] samples visible surfaces of physical objects in a 3D space. In its most basic form, the data resulting from laser scanning is a collection of discrete, densely sampling points in 3D, commonly referred to as a point cloud. A Big Data solution is needed for LiDAR data sets as they are being acquired at increasingly high densities and with greater frequency in many parts of the world including Denmark, England, Finland, Japan, the Netherlands, the Philippines, Slovenia, and Switzerland [1]. The periodic repetition of national and regional LiDAR scans is becoming a more common practice for purposes such as change monitoring. All of these factors contribute to an increasingly significant burden for data storage, management, and processing.

Point cloud data are inherently spatial and share common characteristics with both raster and vector data. However, traditional vector and raster solutions are arguably unsatisfactory for point cloud data storage thereby requiring distinctive data representation strategies [10]. A point cloud data management system is often required to enable access to a large amount of data. A basic example is a data retrieval system that allows users to extract subsets (e.g. using a range search) from within a large point cloud. Data management systems are also frequently used as the backend of point cloud visualization engines. Point cloud subsets need to be fetched from the database for rendering by the visualizer. A range search is also a relevant query in such a scenario. Additionally, point cloud processing systems can be integrated with a supporting database to retrieve the data needed for their processing workflows. Depending on what is needed for the particular processing, different kinds of queries (e.g. range, neighbor, temporal search) may be required. The database, in this case, allows the processing systems to scale to larger datasets.

As set forth by van Oosterom et al. [10], a point cloud data representation should be able to represent the point coordinates (i.e. x, y, z) together with the point attributes (e.g. intensity, color values, classification tags). There should be mechanisms to use spatial coherence to organize the point data, to compress the data, to support concurrent data access by multiple levels-of-detail (LoD), and to control the query accuracy. The authors also suggested a rich set of needed functionalities on point data that include data loading, data querying, simple and complex analysis, data conversions, object reconstruction, LoD use/access, and data updates. Additionally, parallel processing should be considered for all point cloud data operations with the performance of data loading and querying of the utmost importance for a point cloud database implementation. These criteria are the basis for recent developments including the point cloud server by Cura et al. [11], which is a full-fledged, functionality-rich point cloud management system in PostgresSQL built atop the pgPointCloud project.

Given the sheer size of point cloud data being generated and the importance of parallelizing point cloud operations significant research has been undertaken to exploit Big Data approaches for both point cloud analytics and management. They include various applications of tools such as MapReduce and Spark for point cloud processing [12–16]. Such research has proven that generic Big Data analytics frameworks are best-suited for computing problems that are perfectly parallelizable (a.k.a. embarrassingly parallel [17]). Examples include assigning a data point to a raster grid [12, 14] or treating a large point cloud dataset as a group of wholly independent tiles with certain spatial buffer allowance [13]. For computing problems not obviously parallelizable, more complicated strategies such as a master-slave distributed method are needed [15]. Notably, such arrangements may impede the efficiency of the parallelization or even preclude the feasibility of the solution formulation. Research efforts in this area [18–21] are discussed in detail in the next section showing both the state of the art and the general assumption that a Big Data approach will have to be at least part of any future LiDAR point cloud management solution.

2 Related Works on Relational and Non-relational Point Cloud Data Management

This section provides a comprehensive review of major techniques successfully employed for point cloud data management in RDBMS and recent attempts in embracing emerging, distributed, non-relational database technologies. The motivation behind consideration of non-relational databases is also discussed.

In response to the demand for efficient management of spatial data, spatial capabilities have been integrated into a number of RDBMSs including IBM DB2 Spatial Extender, MySQL Spatial, Oracle Spatial, and PostGIS. Some of these spatial DBMSs provide support for point cloud data, often in form of a purpose-built point cloud data representation augmenting the existing spatial capability (e.g. Oracle's SDO_PC, and pgPointCloud's PCPATCH for PostGIS). Without such extensions, generic spatial systems appear to suffer from various performance and storage penalties under significant data volumes [22–24]. The key strategy behind those RDBMS point cloud extensions is the reduction in the indexing granularity. Namely, points are grouped into blocks (a.k.a. chunks or patches), which are handled inside the data system as atomic data units. That reduction significantly decreases the number of data items (e.g. by as little as hundreds to as many as millions of times depending on the specified block size), which in turn decreases the storage, indexing, and management overheads. The drawback of the method is that access to points within a block requires reading and parsing of the entire block. In addition, by grouping points into blocks, certain levels of flexibility in data updating and insertion are lost. Nevertheless, this strategy has been a de-facto standard for point cloud storage in RDBMSs [e.g. 11, 23, 25, 26].

In addition to the aforementioned point grouping method, the use of space filling curves (SFC) is another important strategy increasingly adopted for point cloud storage. A space filling curve is a continuous, surjective mapping from a one-dimensional (1D) space to a higher-dimensional space [27]. Since physical storage devices are essentially 1D and the majority of database systems natively structure data by a singular key, the internal query resolving engine needs to re-formulate multi-dimensional data operations as a 1D problem. SFC is one of the approaches enabling such dimensionality reduction. Thus, it can be exploited to facilitate multidimensional queries on essentially 1D data systems. The use of SFC for point cloud data querying is a specific case of a broader class of data retrieval solutions. Though SFC usage is not restricted to the relational database technology nor point cloud data, there have been many successful attempts to utilize SFC within RDBMSs for point cloud storage and retrieval. Examples include the works by Psomadaki et al. [29], van Oosterom et al. [10]; Vo [26], and Wang and Shan [28]. Interestingly, to store point clouds within an Oracle Index Organized Table, Psomadaki et al. [29] used an SFC, thereby, integrating both space and time as indexes for the data points. Since the SFC-based index already encodes the point coordinates that are selected for indexing (e.g. x, y, z, timestamp), the authors chose not to explicitly store the indexed point coordinates to minimize the storage costs. Even though the storage method allocates one point per row, the method appeared to be highly scalable. That may be attributable to the architecture of the Index Organized Table, which sorted the data by the primary key (i.e. SFC order in this particular implementation). The non-standard architecture is distinguishable from typical RDBMS tables that store the data in their original, unsorted state and maintains separately, rather large data indexes to support the data retrieval.

Even though improvements such as granularity reduction and use of a space filling curve as presented above have made RDBMS point cloud storage viable up to a certain level, there is a certain motivation for considering alternative storage solutions outside the relational database domain. The prime reasons are the demand for greater scalability and performance.

As explained in Sect. 2, Big Data technologies including many non-relational data architectures are not complete replacements for traditional RDBMSs. The prospects of better scalability, performance, availability, and/or functionality of most non-relational data systems come at a cost of lower assurance against violation of traditional database standards such as ACID. The main question while considering a non-relational database implementation is whether some relaxation is tolerable with respect to the specific application requirements. In the context of point cloud data storage, there is an ample space for such compromise. Namely, LiDAR point clouds are often static and rarely require updating. Thus, maintaining data consistency is not as demanding as that for frequently updated data sets in domains such as banking. In addition, point clouds are only weakly relational. Except for the relationship between a point cloud and its metadata, most relationships between point clouds (as well as other geo-spatial datasets) can be implicitly represented via the data's spatial, temporal, and/or spatio-temporal properties. While the potential losses may not be consequential, most of the flexibility provided by non-relational alternatives is meaningful for point cloud management. For example, the schema-less feature allows efficient handling of the heterogeneity of point data derived from different sources. The high level of inherent parallelism, and the potentially high compression rate are amongst the most relevant, favorable traits one can expect from a non-relational database solution.

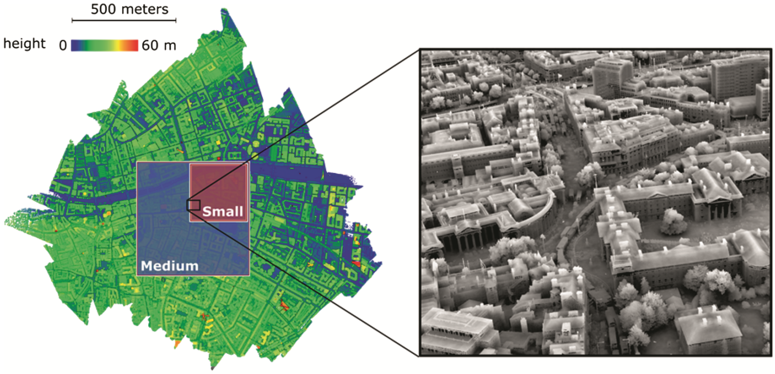
Given the above context, there have been several attempts to employ distributed, non-relational databases for point cloud data management. They include work by Baumann et al. [18], Boehm and Liu [19], Martinez-Rubi et al. [20], and Whitby et al. [21]. For example, selecting MongoDB - a document data store - as its basis, Boehm and Liu [19] stored LiDAR data tiles in their original formats as GridFS files and built a spatial index for the files' spatial extents. That system handles various metadata of the point cloud (e.g. project ID, file type) as BSON (i.e. Binary JSON) documents. Even though this approach is capable of handling a large number of LiDAR files, its usefulness is restricted to file selections since data management is only performed at the file abstraction level. Another investigation of non-relational database for point cloud data management was presented by Martinez-Rubi et al. [20]. In that research, three different approaches for point cloud storage in MonetDB – a column data store – were investigated. All three followed the one point per row method. The first approach indexed point data by the native Imprints indexing in MonetDB, while the second and third approaches used two-dimensional (2D) Morton order (i.e. an SFC approach) to sort the point data. The third differed from the second in the way it replaced the indexed coordinates (i.e. x and y) with the Morton code to achieve a 30% reduction in storage overheads. The authors concluded that SFC-based approaches consumed more time for indexing but were faster and more scalable in querying responses. The authors also emphasized that keeping point data in their binary formats enormously reduced data loading time.

Unlike the works by Boehm and Liu [19] and Martinez-Rubi et al. [20], which are implementations of point cloud storage using existing non-relational databases, EarthServer by Baumann et al. [18] and Geowave by Whitby et al. [21] are full-fledged geospatial data systems capable of accommodating point cloud data. Built atop a multi-dimensional array database architecture, EarthServer is capable of handling and integrating a vast range of Earth observation data types derived from climatic, oceanic, and geological fields for the purpose of spatio-temporal data analytics for data at a petabyte scale. Parallelization across multiple servers is inherently supported at both the inter-query and the intra-query levels (i.e. distributed query processing). As of 2015, point cloud data were experimentally considered as part of the coverage support that also encloses regular grids (i.e. raster), irregular grids, and general meshes. Another system in the same category of integrated geo-spatial database is Geowave [21]. Geowave is an open-sourced, geo-spatial software library capable of augmenting distributed, key-value databases (i.e. Apache Accumulo and Apache HBase) with spatio-temporal functionalities. The main technique backing Geowave is the use of an SFC-based index for row key construction. That technique is the backbone of Geowave, thereby enabling multi-dimensional querying within the key-value data stores. Notably, Geowave is rich in functionality (e.g. data integration, multiple LoD support, MapReduce and Spark integration) and is highly extensible. For example, the concept of Data Adapter (i.e. user-defined data encoder) allows users to model a customized data type of their choice. The point cloud is supported in Geowave via Point cloud Data Abstraction Library (PDAL).

3 Schematic Designs and Implementation of Point Cloud Data Storage in HBase

This paper presents the first steps towards building an integrated distributed, spatio-temporal database system for urban data that takes a point cloud as the central data component. Instead of continuing an existing platform, the authors decided to gather the good practices learned from the existing works to construct several hypothetical point cloud data models atop a representative key-value data store – Apache HBase. The primary purpose is to evaluate the advantages and limitations of each model, as well as to understand the core differences of a non-relational database implementation versus the previous RDBMS works.

To guide the database design and later to aid in evaluating the proposed point cloud data storage models (as described in Sect. 4), an aerial laser scanning dataset of 1.5 km^2 area of the city center of Dublin, Ireland was employed throughout the paper (Fig. 2). The data acquisition was conducted in March 2015, by a Riegl LMS-Q680i scanner. The total number of discrete points captured was 1,420,982,142. The typical local point density on horizontal surfaces was approximately 335 points/m^2 with an approximate vertical surface density about 1/10th of that. The LiDAR point data were delivered in the LAS 1.2 format and occupies approximately 30 GB of disk space. The main data content consisted of a series of point records, in addition to the file-level metadata encoding information such as the spatial extent, the data creation date, and the coordinate system. Each point data record was composed of 28 bytes. The first 12 bytes represent the 3 point coordinates (x, y, z). The subsequent bytes compactly encoded the LiDAR intensity (2 bytes), return number (3 bits), number of returns (3 bits), scan direction flag (1 bit), edge of flight line (1 bit), classification flag (1 byte), scan angle rank (1 byte), user data (1 byte), point source ID (2 bytes), and timestamp (8 bytes). Notably, some attributes can be left blank. Examples include the user data, the point source ID, and the classification attributes.



(a) Full coverage of Dublin LiDAR point cloud (b) 3D rendering of a data subset

Fig. 2. 2015 Dublin point cloud

Given the data structure described in Sect. 1.2, an HBase data model must be constructed from decisions on (1) row-key, (2) column family, (3) column, (4) data cell content, and (5) versioning. Amongst those, the decisions on column family allocation and versioning are relatively obvious, while the others require rigorous evaluation. Since data stored in HBase are physically separated by column family, the column family should be used to group the data that are frequently accessed together. In this implementation, only one column family is allocated given that there is no prior assumption about querying patterns. Regarding the versioning, since the point cloud is static without updating or insertion requirements, only one version is needed. In other words, the versioning function is deactivated presently in this HBase design.

As seen in the literature, representations of a point cloud in binary formats are much more efficient than in text-based formats [10, 23]. As such, the compact LAS encoding is preserved for the point record representation in this paper. The only exception is that whenever possible, empty fields are excluded from storage. The remaining concerns about the data model design are about (1) the row-key design: construct a proper row-key to facilitate the needed queries on the point data; and (2) the column structure: whether all the attributes should be grouped in one column or separated into multiple columns. As there is no insightful reasoning known to the authors, two hypothetical row-key designs (so called Single-Hilbert and Dual-Hilbert) and two column structures (i.e. Separate-Attributes, and Grouped-Attributes) were implemented and experimentally evaluated in this paper. Combinations of these options results in four data models are shown in Fig. 3. Details about the row-key and columns designs are elaborated in Sects. 3.1 and 3.2.

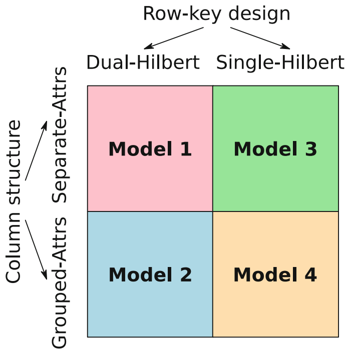


Fig. 3. Conceptual design of four data models for point cloud storage in HBase

3.1 Row-Key Design

Access to data in a key-value storage system is most efficiently performed via data lookup by keys (i.e. row-keys). Consequently, row-key design is the single most important element of any key-value storage solution. A row-key design is driven by dominant data access patterns. Within the scope of this project, 3D spatial queries (i.e. exact point match and range query) were selected as the primary means to access a point cloud data. Given the successful implementations of space filling curves in the authors' previous work [26] as well as in related research [e.g. 20, 21, 29], a 3D Hilbert curve was selected as the primary index to support the specified queries. Hilbert curves are defined to index given 3D spaces (e.g. [4 km × 4 km × 0.5 km]) enclosing the entire spatial extents of the geographical sites of interest at specific resolutions (e.g. 1 m). In the first row-key design named Single-Hilbert, the LiDAR points are indexed by the Hilbert order of the voxel (e.g. [1 m × 1 m × 1 m]) containing the point. All points sharing the same voxel carry the same index. Thus multiple points share the same row-key and are stored on the same row of the database. This multiple-point-per-row method is an approximate analog to the well-founded point cloud storage approaches used in many relational database management systems such as Oracle's SDO_PC and pgPointCloud's PCPATCH.

To evaluate the suitability of the aforementioned traditional solution relative to the simple and more intuitive one-point-per-row approach (a.k.a. flat model), a second row-design called Dual-Hilbert was developed. With the Dual-Hilbert solution, the 1-m resolution Hilbert index (i.e. coarse Hilbert index) is concatenated with a second spatial index computed locally within the voxel at a finer resolution (e.g. 1 mm) [i.e. local Hilbert index]. The fine resolution is set to be finer than the point data resolution (i.e.

centimeter range) so that the concatenated Hilbert code is unique for each point. This flat model approach abandoned by the relational database systems has the potential to surpass more traditional multiple-point-per-row approaches since vertical databases such as HBase perform best for "tall tables" (i.e. large number of rows with small amount of data stored per row). Another potential benefit of using the fine-grained Hilbert index is that it can be used as a more compact replacement for the point coordinates to reduce total disk and network I/O costs. The comparisons of Model 1 versus Model 3, and Model 2 versus Model 4 (Fig. 3) aim to provide evidence to assess these hypotheses.

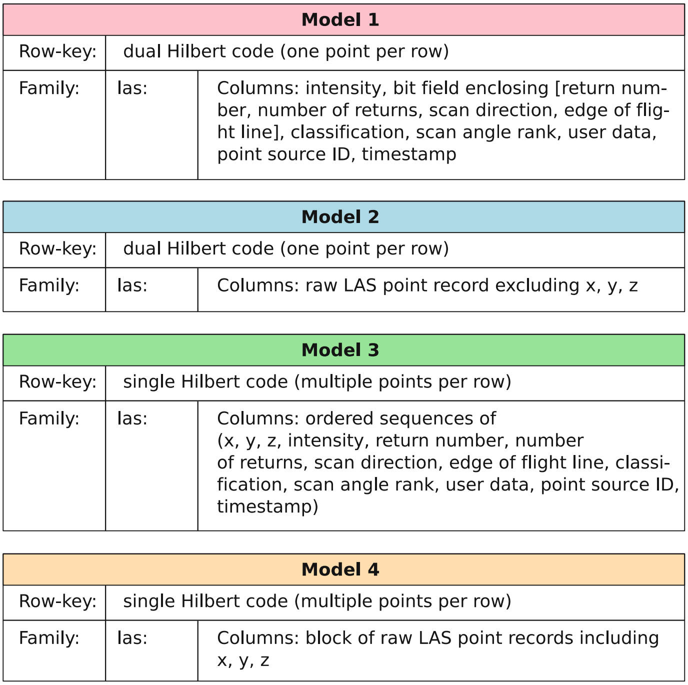
3.2 Column Structures

As mentioned previously, each LiDAR data point contains a range of attributes in addition to the point coordinates. There are 10 such attributes for each point in the Dublin scan; some of which contain empty values. However, the structure and the number of point attributes are not the same for every LiDAR scan. File-based approaches including the widely used LAS data exchange format handle that semi-structured situation by providing a set of fixed templates to cover common data patterns. Each template contains a pre-determined set of placeholders for point attributes. Users can then choose the template that best matches their data, amongst the limited choices offered by the format specification. This lack of versatility can be completely alleviated when column family data stores such as HBase are used, as they are schema-less. Column family stores have no restriction on data type, data name, or the number of the attributes stored in each row. Two rows in the same table can have completely different sets of attributes. However, the flexibility and versatility do not come without a cost (e.g. storing the attribute names for each value). To better analyze the gains and costs of using HBase to provide a flexible point attribute structure, two column structures are investigated in this study. The first structure (i.e. Separate-Attributes) separates the point attributes into different columns so that the points in a table do not have to conform to any fixed template. The second structure (i.e. Grouped-Attributes) assimilates the LAS approach, in which all attributes are maintained as a binary array in a fixed structure. Comparisons of Model 1 versus Model 2, and Model 3 versus Model 4 aims to provide evidence for the assessment of column structures as shown in Sect. 4.

3.3 Query Processing

This section presents the query resolving strategies with respect to the four data models constructed in Sects. 3.1 and 3.2, which are depicted in a less abstract way in Fig. 4. Point query (a.k.a. exact point match) is the most basic type of query on point cloud data. Point query aims to search for an exact match of a given point [i.e. an (x, y, z) triplet] and return all the associated attributes of the found point record. For the Dual-Hilbert data models, the exact match can be directly computed by transforming the (x, y, z) triplet into a dual Hilbert code, and then looking up the corresponding HBase table for a row-key matching the Hilbert code. Such lookup by key in HBase is termed `Get`.

Resolving point queries for Single-Hilbert models is slightly more complicated and less efficient. Since only the coarse Hilbert code (1 m resolution) is available for lookup,



Model 1		
Row-key:	dual Hilbert code (one point per row)	
Family:	las:	Columns: intensity, bit field enclosing [return number, number of returns, scan direction, edge of flight line], classification, scan angle rank, user data, point source ID, timestamp

Model 2		
Row-key:	dual Hilbert code (one point per row)	
Family:	las:	Columns: raw LAS point record excluding x, y, z

Model 3		
Row-key:	single Hilbert code (multiple points per row)	
Family:	las:	Columns: ordered sequences of (x, y, z, intensity, return number, number of returns, scan direction, edge of flight line, classification, scan angle rank, user data, point source ID, timestamp)

Model 4		
Row-key:	single Hilbert code (multiple points per row)	
Family:	las:	Columns: block of raw LAS point records including x, y, z

Fig. 4. Detail data schema of the four data models

the returned data from a `Get` function is a block of points, which must be parsed and validated to return the ultimate querying result. The point block parsing can be done at either the client side or the server side in HBase. A server-side implementation (e.g. a HBase Custom Filterer) is more complicated but is more efficient as it avoids sending the entire block of points through the network, and the computing power on the server side is more available to handle the computation. Ultimately, from the theoretical perspective, a design that separates attributes into distinct columns (i.e. Model 1 and Model 3) should result in better querying performance in cases where only a subset of the attributes is requested.

Spatial range search is another important type of point cloud query. A range query returns all data points enclosed within a given querying window, which often has the form of a 3D polygonal shape. The simplest case of a querying window is a rectilinear box. Spatial range search is useful for applications including downloading a data subset or clipping point data by a viewing frustum for visualization. Resolving a range query on point data sorted by a space filling curve involves decomposing the bounding of the querying windows into several continuous Hilbert segments (i.e. 1D numeric segments). The number of Hilbert segments is equivalent to the number of 1D range searches invoked against the database. Data querying can be slow, if a querying window is highly fragmented and requires a large number of range searches. The fragmentation issue can be alleviated by loosening the continuous constraint within each Hilbert segment. Namely, if the separation between two Hilbert segments is smaller than a certain level (e.g. 500 cells), the segments are grouped to reduce the number of total segments (i.e. number of database invokes). The strategy can greatly accelerate data queries at the cost of including more false-positive results (i.e. the gaps within the Hilbert segments), which may result in higher pressure on later filtering steps. Setting the value for the allowable Hilbert gap to optimize query performance is the matter of balancing the two factors and requires empirical tuning.

The Hilbert segments are then used to retrieve the point candidates by the native 1D range search on the row-keys. The Hilbert order only facilitates a coarse filtering for range querying. Namely, the candidate points resulting from a Hilbert decomposition include not only the true result but also some false positive points. The false positive points include those that fall outside the querying window but are inside its bounding box or those share the same Hilbert order with the actual resulting points. In order to get the exact results, a final fine-filtering is needed to perform a spatial check for each and every candidate points returned by the Hilbert coarse filtering. This relatively costly fine filtering should preferably be pushed to the server side to take advantage of the parallelism and data locality. Compared to the Single-Hilbert models, the dual level Hilbert codes in Model 1 and Model 2 allow Hilbert filtering at one extra level, thus reducing the amount of data passed through the fine filtering. The fine filtering can be skipped for some applications that can tolerate false positive points such as many visualizations. Skipping the fine filtering can greatly accelerate querying speed and is done natively in some existing systems such as pgPointCloud [23]. In the current HBase implementations introduced in this paper, the fine filtering can be enabled or disabled at the time of querying.

3.4 Data Ingestion and Querying Workflows

Starting with a large, unstructured point data set in the binary LAS format, this section introduces a workflow for loading the data into HBase, while exploiting the Hadoop distributed framework (Fig. 5). One of the issues during the data ingestion is that the original data format (i.e. LAS) is not suitable for processing in parallel on a computing cluster. To address that, the original LAS formatted data are transformed into a Hadoop Sequence File format (Step 1 in Fig. 5). Sequence File (SF) is a Hadoop mechanism for encapsulating arbitrary binary data into a key-value format, while making the data splittable for parallel processing on Hadoop clusters. The LAS-to-Sequence transformation parses point records from the input LAS files and encodes them as values in the corresponding SFs. In this particular case, the point cloud SF's keys are not useful and, thus, left blank. The SF's metadata, which is a set of text-based key-value pairs, is exploited to store offset and scale parameters; the information needed for parsing the LAS point data. Subsequent to the sequential transformation, which occurs outside the cluster, the point cloud is uploaded to the Hadoop Distributed File System (HDFS) (Step 2 in Fig. 5). Thus, the transformation to Sequence File enables parallel processing of large point cloud data.

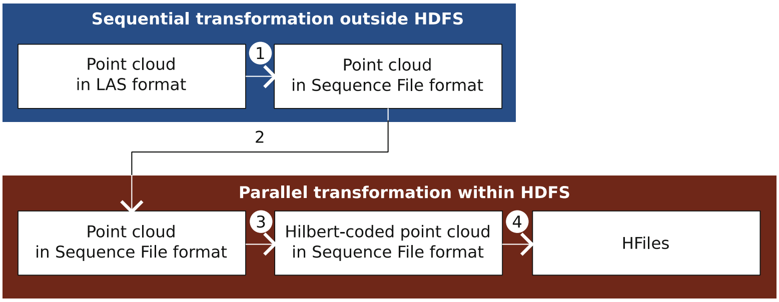

Sequential transformation outside HDFS
Point cloud in LAS format
1
Point cloud in Sequence File format
2
Parallel transformation within HDFS
Point cloud in Sequence File format
3
Hilbert-coded point cloud in Sequence File format
4
HFiles

Fig. 5. Data ingestion workflow

The data ingestion procedure continues with the Hilbert computation for every point record. A MapReduce program is used for this purpose (Step 3 in Fig. 5). The mapper computes a coarse Hilbert code for each point record and outputs the result as `<coarse-hilbert; raw-las-point-record>`. The sort and shuffle process automatically handled by Hadoop MapReduce is responsible for grouping the point records by the coarse Hilbert code. Arriving at the reduce phase, the point data are grouped into blocks by the coarse Hilbert codes. The fine Hilbert codes are then computed, and the resulting codes are appended ahead of each point of the block. Ultimately, the Sequence Files resulting from the Hilbert computation have the format of `<coarse-hilbert; fine-hilbert-coded-point-block>`. Prior to being loaded into HBase, the point data need to be transformed once more into a so-called HFile format, which is the native file format underlying HBase (Step 4 in Fig. 5). During this transformation, the row-key, column family, column, and data content are set. The final step of ingesting HFile data into HBase tables and distributing the data across multiple servers is facilitated by an in-built HBase function.

4 Performance Evaluation

To aid in the performance evaluation, three different subsets of varying sizes (Small - S, Medium - M, and Large - L) were extracted from the 2015 Dublin point cloud. The Large dataset includes all the 2015 Dublin point cloud while the coverages of the Small and Medium datasets are shown in Fig. 2. The approximate number of points in the S, M and L datasets are 90 million, 360 million, and 1.43 billion, respectively. All of the experiments were run on a 10-node Hadoop cluster. Each node consists of 2 Intel Haswell (E5-2695 v3) CPUs and 128 GB DDR4-2133 RAM.

4.1 Data Ingestion

Amongst the 4 steps of the data ingestion process presented in Fig. 5, only Step 4 is model-dependent and needs to be run for each of the 4 models. The results of the other 3 steps are shared for all data models. Thus, they were executed only once. The LAS-to-Sequence transformation processed the data with multi-threading parallelization at a speed of 2,855,020 points/sec. The Hilbert computation operated on the cluster at a speed of 813,536 points/sec. The HFile creation speed, total data loading speed, and disk consumption of the all the experiments corresponding to the 3 datasets and the 4 models are presented in Table 1.

Table 1. Data loading speed and disk consumption

Data model	Dataset	Number of points	Size (bytes/point)	HFile creation speed (points/sec)
1	S	89,970,106	244.5	57,285
2	S	89,970,106	51.3	182,324
3	S	89,970,106	32.2	944,593
4	S	89,970,106	28.4	1,250,587
pgpc	S	89,970,106	21.0	52,698[a]
1	M	365,612,527	244.5	60,914
2	M	365,612,527	51.3	177,407
3	M	365,612,527	32.0	1,328,281
4	M	365,612,527	28.4	1,908,530
pgpc	M	365,612,527	21.0	61,939[a]
1	L	1,420,982,142	235.0	41,283
2	L	1,420,982,142	48.3	181,110
3	L	1,420,982,142	31.2	1,344,047
4	L	1,420,982,142	26.9	2,372,243
pgpc	L	1,420,982,142	21.0	57,091[a]

[a]Total data ingestion time into a pgPointCloud database using PDAL (https://www.pdal.io)

The corresponding storage costs and data loading speeds of the equivalent tests using a pgPointCloud database [23] are also included in the table for comparison. Except for Model 1, the total ingestion time for all other 3 data models, including the Hadoop Sequence file conversion and Hilbert computation, were significantly shorter [i.e. 2.5 to 8.0 times] than the time needed to load data into pgPointCloud databases. The data ingestion times for Model 1 were 1.5 times longer than that for pgPointCloud. The disk consumption of all the HBase models was higher than the pgPointCloud data.

All data models including the pgPointCloud databases appear scalable, as the data speed remains relatively constant with respect to significant increases in data size. The disk consumption appears to largely independent of the total data size. The HBase models can be sorted as Model 1, Model 2, Model 3, Model 4 in a descending order of both the data loading speed and the storage costs. Model 1, which stores one point per row with point attributes separated in different columns, is significantly larger and took much more time to load. The above experimental results can be interpreted as the separation of point attributes, and the use of dual Hilbert codes introduces significant overheads, which is understandable when considering the physical storage structure in HBase. Namely, HBase stores data as key-value pairs. While the total amount of values – the actual point data content - is unchanged among the models, the keys vary largely. A key in HBase is an aggregation of row-key, column family name, column name, and timestamp. Both the separation of attributes and the use of dual Hilbert code increase the number of key-value pairs. In addition, the former results in significantly more content stored in the keys. The empirical results show that the flexibility in the data schema provided by HBase as maximized in the Separate-Attributed data models comes with significant overheads.

4.2 Point Query

One thousand points subsampled from the Small dataset were used as querying points for the point query performance evaluation. The queries were executed consecutively. The first query was a cold query, which often takes a longer time to process compared to subsequent queries (also known as hot queries). The difference between hot and cold query speed is mostly attributable to caching.

The point query response times are presented in Fig. 6 (hot queries), Fig. 7 (cold queries), and Table 2. Box plots are selected to present the distribution of the performance of the 1000 hot queries. In each box plot, the notched rectangle (i.e. box) represents the middle 50% of the response time values. The bottom and the top of the box are called the lower and the upper quartiles, which bound the middle 50% of the samples. The notch itself shows the median value, which equally bisects the sample population. The

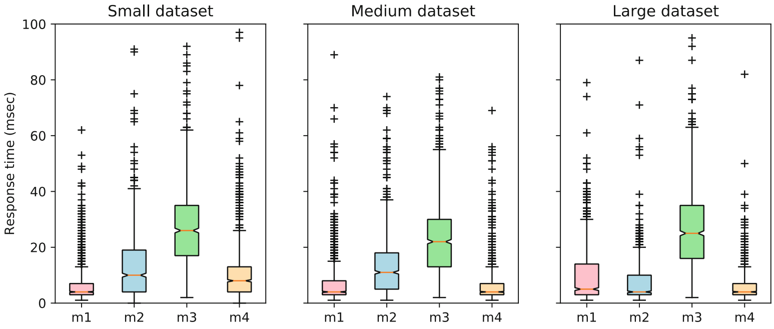


Fig. 6. Hot point query response times

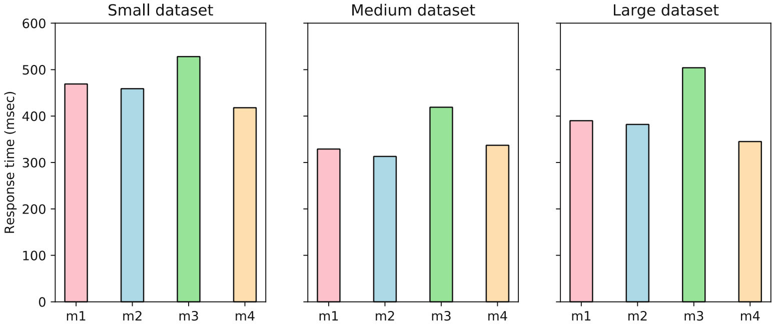


Fig. 7. Cold point query response times

Table 2. Point query response time

Data model	Dataset	Point query response time (msec)	
		Hot (median)	Cold
1	S	4	469
2	S	10	459
3	S	26	528
4	S	8	418
pgpc	S	124	120
1	M	4	329
2	M	11	323
3	M	22	419
4	M	4	337
pgpc	M	135	130
1	L	5	390
2	L	4	382
3	L	25	504
4	L	4	345
pgpc	L	134	110

crosses are outliers, which are the values exceeding 1.5 times the upper quartiles or lesser than 1.5 times the lower quartiles. The two whiskers projected from the box represents the values outside the middle 50% of the population excluding the outliers.

As expected, the cold queries were approximately 300 to 500 ms slower than the hot queries. For both cold and hot queries, Model 3 appeared to be the slowest requiring 400 to 500 ms for the first runs and 22 to 26 ms for the subsequent invokes. For the other models, cold queries were in the range of 300 to 450 s, while hot queries were from 4 to 11 s. The lower performance of Model 3 was caused by the difference between the stored data structure and point record structure returning to the queries. More specifically, as seen in Fig. 4, point data stored in Model 3 are partitioned by attributes. Each cell contains an ordered sequence of the same attributes, e.g. x_0, x_1, x_2, …, x_n. In response to a query, these attribute sequences need to be parsed and re-organized into the point record structure, e.g. x_0, y_0, z_0, $intensity_0$, …, $timestamp_0$. The restructuring overhead is the reason for the lower performance of Model 3. Nevertheless, the most significant observation extracted from the experiment demonstrated that all data models were scalable in point querying. There was no observable performance degradation with respect with the growth in data volume.

Table 2 also presents the corresponding querying response times for point data stored in a PostgreSQL database with a pgPointCloud extension [23]. Since pgPointCloud does not support exact match queries, approximately equivalent range queries were used with additional filtering by the Point Data Abstraction Library (PDAL). The querying scripts can be seen in Appendix B. Compared to Model 3 – the slowest model amongst the 4 HBase candidates – exact match queries with pgPointCloud were at least 5 times slower. The differences between the pgPointCloud performance and the other 3 data models were from 12 to 34 times. Notably, there was no observable difference between the cold and hot queries in the pgPointCloud tests.

4.3 Range Query

The first 50 samples of the querying points used for testing the point queries in Sect. 4.2 were re-used to evaluate the performance and scalability of the 4 data models in supporting range queries. Two classes of range queries were investigated. The first type considered small querying windows that are cubes with 3 m long sides. The second class considered large querying windows having a side length of 50 m. To alleviate the effects of data density, the range query response times were normalized by the number of returning points. The results of cold queries and hot queries corresponding to each class are plotted in Figs. 8, 9, 10, 11, and Table 3. Equivalent tests of PostgreSQL pgPointCloud databases are reported alongside the tests of the 4 data models for comparison. Notably, the HBase and pgPointCloud tests are not completely equivalent, because pgPointCloud only supports 2D queries, while the HBase queries are in 3D. All the response times reported in this section include the entire costs for extracting data from the databases and exporting the resulting points to LAS files.

An important observation is that all 4 data models and the pgPointCloud databases are perfectly scalable. The querying response times remained largely unchanged, despite the growth in data volume. Model 4 appears to be the best performer amongst the investigated solutions. Model 4 was 3 to 4 times faster than the other 3 data models in the cases of small, hot queries. The factors were larger (i.e. from 4 to 8 times) for the large queries. Compared to pgPointCloud, Model 4 was consistently faster (i.e. from 2 to 4 times). The difference between Model 4 and pgPointCloud was less significant in the tests with the large querying windows. The better performance of Model 4 compared to the other 3 data models is likely to be attributable to several factors. First, the aggregation of points into blocks and the attribute grouping greatly reduced the number of key-value pairs. Despite the side-effect of having larger datum per value, the number of key-value pairs reduction seems to have had a positive effect on both the querying time and the storage overhead. The second factor contributing to the better performance of Model 4 was that the data model preserved the original structure of LAS point data records, which was also the data format returned to the range queries. As such, the binary sequences stored in the database were returned directly without having to undergo restructuring as was required in the other models.

Similar to what was observed from the point queries, Model 3 was also the slowest amongst all the solutions with respect to range queries. In fact, Model 1, 2, and 3 were all slower than pgPointCloud, which has an underlying structure similar to Model 4. More specifically, the point data in both pgPointCloud and Model 4 were grouped into spatial coherent groups, while the attributes of each point record were serialized into a fixed-length binary string. The observation demonstrated that the concepts established for enhancing the performance and scalability of point cloud storage in traditional environments are also applicable to distributed databases.

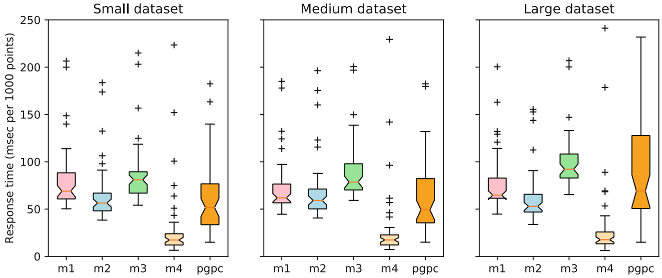


Fig. 8. Hot range query response times for small querying windows

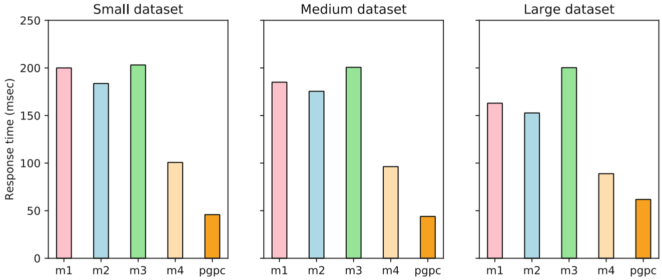


Fig. 9. Cold range query response times for small querying windows

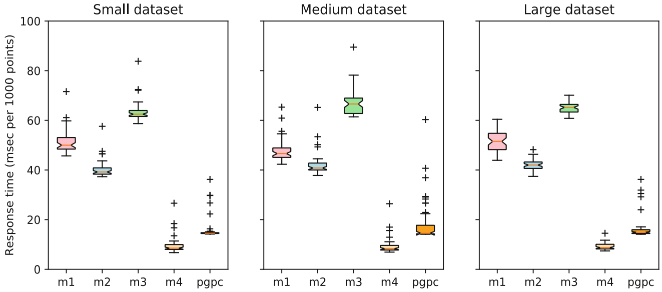


Fig. 10. Hot range query response times for large querying windows

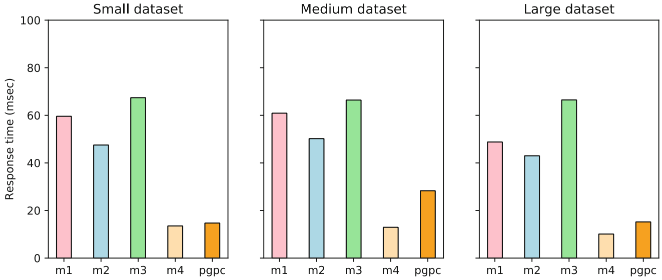


Fig. 11. Cold range query response times for large querying windows

Table 3. Range query response times

Data model	Dataset	Point query response time (msec per 1000 points)			
		Small queries [3 × 3 × 3]		Large queries [50 × 50 × 50]	
		Hot	Cold	Hot	Cold
1	S	69	200	50	60
2	S	56	184	39	48
3	S	81	203	62	67
4	S	17	101	9	14
pgpc	S	52	46	14	15
1	M	62	185	47	61
2	M	59	175	41	50
3	M	79	201	67	66
4	M	17	96	8	13
pgpc	M	49	44	15	28
1	L	65	163	51	49
2	L	53	153	42	43
3	L	92	200	65	67
4	L	18	89	9	10
pgpc	L	69	62	15	15

For all 4 models and pgPointCloud, the unit querying costs per point decreased with larger querying windows. A more detailed analysis actually shows that the decrease stops after the querying size reaches to a certain level (e.g. around 30 m in the investigated tests). This may be due to some overheads that are independent of the number of resulting points. When more points are returned from larger querying windows, the distribution of the overheads per point gets smaller and becomes insignificant at a certain querying size. The same logic is behind the dissimilarity of the hot and the cold query response times in the cases of large querying windows. In these large queries, the overheads needed in the first query get distributed to more points and become insignificant fractions. Notably, the cold queries of pgPointCloud databases does not appear to be slower than subsequent queries. Thus, the first queries of pgPointCloud were sometimes faster than the corresponding queries of all the HBase data models.

5 Concluding Remarks

As a demonstration for an implementation of a distributed, non-relational, key-value store for large and high-resolution point cloud data, this paper presents four data models for storage, indexing, and querying point clouds. The four models are constructed from two row-key designs (i.e. Single-Hilbert and Dual-Hilbert) and two column structures (i.e. Separate-Attributes and Grouped-Attributes). The Dual-Hilbert models resemble the flat model approach in RDBMS point cloud storage, while the Single-Hilbert models are largely similar to the standard point block solution. In addition, the Dual-Hilbert codes were used as replacement for the point coordinates. The experimental evaluations of up to 1.4 billion points showed that the flat models are as scalable as the block models within HBase, unlike what has been observed in traditional RDBMS environments. The only notable demerit of the flat models is that they required more storage space and were slower to create initially, without any benefit in the querying speed. The two columns structures, Separate-Attributes and Grouped-Attributes, were compared to evaluate the capability of HBase in supporting flexible data schema. The separation of point attributes to different columns allowed the heterogeneity in point record structure and avoided storage of empty fields. However, doing so in HBase resulted in significant storage overhead as reflected by the sharp increase in the number of key-value pairs and the longer key content. That increase in the number of stored data entities seemed to affect the querying performance in the case of range querying.

Amongst the investigated data models, Model 4, which indexes point data at the block level and preserves the aggregation of the point attributes, appears to be the most competitive solution. The simple structure of Model 4 allows the data to be loaded 7 to 46 times faster than the Dual-Hilbert models (Model 1, Model 2) and at least 1.3 times faster than Model 3. Range queries with Model 4 are from 3 to 8 times faster than the other models, while its point query performance is among the highest. Future research will investigate Model 4 further with regard to its capability to support queries that seek only a subset of point attributes. Since Model 4 does not index the data at the point attribute level as in Model 1 and Model 3, there is a potential that it may not be as effective as the other models in supporting the attribute-specific queries. In addition, heterogeneous datasets (i.e. point data with various attribute structures) will be used to further evaluate the storage efficiency of the data models and explore the schema-less feature of HBase.

The evaluation against pgPointCloud, which is an existing relational database solution, showed that all the HBase data models were faster than pgPointCloud in supporting point queries. With respect to range queries, Model 4 was from 1.5 to 4 times faster than pgPointCloud. However, the other 3 HBase data models were slower than the traditional solution. The result shows that grouping data into blocks and preserving the point record structure are good strategies for encoding point cloud data in HBase. Notably, the unit querying speed per point of the range queries decreased with a larger querying size. The differences between the first (i.e. cold) queries and subsequent (i.e. hot) queries were also reduced when the query size expanded. Due to the built-in parallel mechanism of Hadoop, loading point data into HBase was considerably faster than the pgPointCloud data ingestion despite the requirement of some data preprocessing steps. There was an exception with Model 1 where the extreme indexing decelerated the data ingestion to as much as 1.5 times slower than the pgPointCloud. Finally, these advances do come at a cost. Namely, all the HBase data models consumed more disk space than the pgPointCloud.

In summary, distributed, non-relational databases can be promising for point cloud data storage, because point clouds are weakly relational and do not strictly require transactional consistency. The most significant gains expected from migrating to a non-relational alternative include an improved possibility to scale the system for large amounts of data and better performance due to the inherent parallelism in the framework. The experimental results presented in this paper show that HBase, a representative distributed database, was scalable and faster than the relational PostgreSQL pgPointCloud database when similar data encoding strategies were used (Model 4). The storage of one point per row in HBase (Model 1 and Model 2) did not encounter a scalability issue as previously observed in relational databases [10]. However, they were slower than the storage scheme that groups data into blocks. Future research should consider different techniques to further optimize the performance of both the non-relational and relational solutions. Testing the databases with data of greater volumes and complexity should also be considered.

Acknowledgments. The Hadoop cluster used for the work presented in this paper was provided by allocation TG-CIE170036 - Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 [30]. The authors would like to thank the staff at Pittsburg Supercomputing Center for the truly outstanding technical support provided during setting up the testing. This research also made use of data collected with funding from the European Research Council grant ERC-2012-StG 20111012 “RETURN - Rethinking Tunnelling in Urban Neighbourhoods” Project 307836.

The dataset is available from NYU Spatial Data Repository https://doi.org/10.17609/N8MQ0N.

References

1. Vo, A.V., Laefer, D.F., Bertolotto, M.: Airborne laser scanning data storage and indexing: state of the art review. Int. J. Remote Sens. **37**(24), 6187–6204 (2016). https://doi.org/10.1080/01431161.2016.1256511
2. Kitchin, R., McArdle, G.: What makes Big Data, Big Data? exploring the ontological characteristics of 26 datasets. Big Data Soc. **3**(1), 1–10 (2016). https://doi.org/10.1177/2053951716631130
3. Ghemawat, S., Gobioff, H., Leung, S.T.: The Google file system. In: Proceedings of the 19th ACM Symposium Operating Systems Principles, New York, pp. 29–43 (2003)
4. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2004). https://doi.org/10.1145/1327452.1327492
5. White, T.: Hadoop The Definitive Guide, 4th ed. O’Reilly, Massachusetts (2015)
6. Chang, F., et al.: Bigtable: a distributed storage system for structured data. ACM Trans. Comput. Syst. **26**(2), 4 (2008). https://doi.org/10.1145/1365815.1365816
7. George, L.: HBase The Definitive Guide, 1st edn. O’Reilly, Massachusetts (2011)
8. Middleton, W., Spilhaus, A.: The measurement of atmospheric humidity. In: Meteorological Instruments, Toronto, pp. 105–111 (1953)
9. Shepherd, E.C.: Laser to watch height: New Scientist, vol. 6, no. 437, p. 33 (1965)
10. van Oosterom, P., et al.: Massive point cloud data management: design, implementation and execution of a point cloud benchmark. Comput. & Graph. **49**, 92–125 (2015). https://doi.org/10.1016/j.cag.2015.01.007
11. Cura, R., Perret, J., Paparoditis, N.: A scalable and multi-purpose point cloud server (PCS) for easier and faster point cloud data management and processing. ISPRS J. Photogramm. Remote Sens. **127**, 39–56 (2017). https://doi.org/10.1016/j.isprsjprs.2016.06.012
12. Krishnan, S., Baru, C., Crosby, C.: Evaluation of MapReduce for gridding LIDAR data. In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science, pp. 33–40 (2010). https://doi.org/10.1109/cloudcom.2010.34

13. Li, Z., Hodgson, M.E., Li, W.: A general-purpose framework for parallel processing of large-scale LiDAR data, vol. 8947. Int. J. Digit Earth **11**(1), 26–47 (2017). https://doi.org/10.1080/17538947.2016.1269842
14. Rizki, P.N.M., Eum, J., Lee, H., Oh, S.: Spark-based in-memory DEM creation from 3D LiDAR point clouds. Remote Sens. Lett. **8**(4), 360–369 (2017). https://doi.org/10.1080/2150704X.2016.1275053
15. Hamraz, H., Contreras, M.A., Zhang, J.: A scalable approach for tree segmentation within small-footprint Airborne LiDAR data. Comput. Geosci. **8**(4), 360–369 (2017). https://doi.org/10.1080/2150704X.2016.1275053
16. Aljumaily, H., Laefer, D.F., Cuadra, D.: Urban point cloud mining based on density clustering and MapReduce. J. Comput. Civ. Eng. **31**(5) (2017). https://doi.org/10.1061/(asce)cp.1943-5487.0000674
17. Moler, C.: Matrix computation on distributed memory multiprocessors. In: Hypercube Multiprocessors 1986, pp. 181–195 (1987)
18. Baumann, P., et al.: Big Data analytics for Earth sciences: the EarthServer approach. Int. J. Digit. Earth **9**(1), 3–29 (2015). https://doi.org/10.1080/17538947.2014.1003106
19. Boehm, J., Liu, K.: NoSQL for storage and retrieval of large LiDAR data collections. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, **XL-3**/W3, pp. 577–582, La Grande Motte (2015)
20. Martinez-Rubi, O., et al.: Benchmarking and improving point cloud data management in MonetDB. SIGSPATIAL Special - Big Spatial **6**(2), 11–18 (2014). https://doi.org/10.1145/2744700.2744702
21. Gertz, M., Renz, M., Zhou, X., Hoel, E., Ku, W.-S., Voisard, A., Zhang, C., Chen, H., Tang, L., Huang, Y., Lu, C.-T., Ravada, S. (eds.): SSTD 2017. LNCS, vol. 10411. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64367-0
22. Mosa, A.S.M., Schön, B., Bertolotto, M., Laefer, D.F.: Evaluating the benefits of octree-based indexing for LiDAR data. Photogramm. Eng. Remote Sens. **78**(9), 927–934 (2012). https://doi.org/10.14358/PERS.78.9.927
23. Ramsey, P.: LiDAR in PostgreSQL with PointCloud. In: FOSS4G, Nottingham (2013)
24. Nandigam, V., Baru, C., Crosby, C.: Database design for high-resolution LIDAR topography data. In: Gertz, M., Ludäscher, B. (eds.) SSDBM 2010. LNCS, vol. 6187, pp. 151–159. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13818-8_12
25. Murray, C., et al.: Oracle Spatial and Graph - developer' s guide, 12c Release 1 (2017). https://docs.oracle.com/database/121/SPATL/toc.htm
26. Vo, A.-V.: Spatial data storage and processing strategies for urban laser scanning. Ph.D. thesis. University College Dublin (2017). https://doi.org/10.13140/rg.2.2.12798.48962
27. Haverkort, H., van Walderveen, F.: Locality and bounding-box quality of two-dimensional space-filling curves. Comput. Geom. **43**(2), 131–147 (2008). https://doi.org/10.1016/j.comgeo.2009.06.002
28. Wang, J., Shan, J.: Space-filling curve based point clouds index. In: Proceedings of the 8th International Conference on GeoComputation, Michigan (2005)
29. Psomadaki, S., van Oosterom, P.J.M., Tijssen, T.P.M., Baart, F.: Using a space filling curve approach for the management of dynamic point clouds. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2/W1, pp. 107–118 (2016). https://doi.org/10.5194/isprs-annals-iv-2-w1-107-2016
30. Towns J., Cockerill T., Dahan M., Foster I., Gaither K., Grimshaw A., Hazlewood V., Lathrop S., Lifka D., Peterson G.D., Roskies R., Scott J.R., Wilkins-Diehr N.: XSEDE: accelerating scientific discovery. Comput. Sci. Eng. **16**(5), 62–74 (2014). https://doi.org/10.1109/mcse.2014.80

SLAM-Driven Intelligent Autonomous Mobile Robot Navigation for Construction Applications

Pileun Kim, Jingdao Chen, Jitae Kim, and Yong K. Cho

Georgia Institute of Technology, Atlanta, GA 30332, USA
{pkim45,jchen490,jkim3181}@gatech.edu
yong.cho@ce.gatech.edu

Abstract. The demand for construction site automation with mobile robots is increasing due to its advantages in potential cost-saving, productivity, and safety. To be realistically deployed in construction sites, mobile robots must be capable of navigating in unstructured and cluttered environments. Furthermore, mobile robots should recognize both static and dynamic obstacles to determine drivable paths. However, existing robot navigation methods are not suitable for construction applications due to the challenging environmental conditions in construction sites. This study introduces an autonomous as-is 3D spatial data collection and perception method for mobile robots specifically aimed for construction job sites with many spatial uncertainties. The proposed Simultaneous Localization and Mapping (SLAM)-based navigation and object recognition methods were implemented and tested with a custom-designed mobile robot platform, Ground Robot for Mapping Infrastructure (GRoMI), which uses multiple laser scanners and a camera to sense and build a 3D environment map. Since SLAM did not detect uneven surface conditions and spatiotemporal objects on the ground, an obstacle detection algorithm was developed to recognize and avoid obstacles and the highly uneven terrain in real time. Given the 3D real-time scan map generated by 3D laser scanners, a path-finding algorithm was developed for autonomous navigation in an unknown environment with obstacles. Overall, the 3D color-mapped point clouds of construction sites generated by GRoMI were of sufficient quality to be used for many construction management applications such as construction progress monitoring, safety hazard identification, and defect detection.

Keywords: Mobile robot · Navigation · SLAM · Object recognition
Unstructured environment · Construction · Point clouds

1 Introduction

The current commercial 3D laser scanning solutions for the construction industries can collect millions of three-dimensional points in a short period of time accurately and safely under stationary conditions. However, the post-processing, which is point cloud registration process to combine each point cloud from different scan locations into one coordinate system, is still a labor-intensive and time-consuming process. First, multiple scans from different scan locations are manually gathered by the operator with targets which are marked correspondences in the scan area to find common points between point

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 254–269, 2018.
https://doi.org/10.1007/978-3-319-91635-4_14

clouds. Second, it can cause errors in placing and relocating targets. Third, these collected data are registered with matching common features or targets manually. However, the operator does not know until the end of the registration process that the collected data contain imperfections due to incomplete scans, hindering structures, absence of targets and lack of common features for registration in many cases. This is because current laser scanning methods provide limited feedback to the operator during the scan process and registration process [1]. As like this, the current quality control methods at a job site rely heavily on manual inspection which is labor-intensive, tedious, and error-prone. Also, the raw data collected from construction sites are often unstructured and poorly organized which results in time and cost overheads in interpreting the data as well as delays in sharing the relevant information to stakeholders [2]. Therefore, robot technology has the potential to reduce construction and maintenance costs and improve productivity, quality, and safety. The automated data collection of construction site conditions by a mobile robot would provide an attractive alternative for the execution of routine work tasks at a job site [3, 4].

To be a practical solution for construction applications, an autonomous mobile robot must have the capability to safely move around an unstructured and cluttered environment with many spatiotemporal objects and obstacles which continuously change the environment conditions (i.e., uneven ground surface, equipment, materials, soil stockpiles, temporary structures). Robot's obstacle recognition and avoidance mechanisms should be fast, robust and not solely dependent on the as-designed construction information of the environment because quickly changing as-is conditions of a construction job site are different from as-designed conditions in most cases. Several methods exist in the literature for robot navigation in other domains which work by observing an existing map and planning the robot's every move in advance [5, 6]. However, the existing methods are not suitable for construction applications due to the aforementioned construction site's unique characteristics.

This paper introduces an autonomous as-is 3D spatial data collection method by a mobile robot specifically aimed for construction job sites with many spatial uncertainties. The mobile robot system uses a simultaneous localization and mapping (SLAM) techniques to determine its navigation paths and construct a 3D point cloud of a construction site. While SLAM provides a robot's current location, it does not detect obstacles or uneven surface conditions. The kinematic modeling of the mobile robot was analyzed, and a fuzzy control was developed for robot's autonomous navigation in an unknown environment with obstacles. Additionally, obstacle detection algorithms were used to recognize spatiotemporal obstacles to avoid in real time. The proposed SLAM and object-recognition methods were implemented and tested with a custom-designed mobile robot platform, Ground Robot for Mapping Infrastructure (GRoMI), which uses multiple laser scanners along with infrared and sonar sensors to sense and build a 3D environment map. To evaluate the overall navigation and object recognition performance of GRoMI, two experiments were conducted: one for an indoor building site and the other for an outdoor construction site. The following sections will present an overview of the related work, methodology, results, conclusion, and discussion.

2 Literature Review

2.1 Mobile Robot Navigation

Path planning in static and dynamic unknown environments had been receiving considerable attention over the past years due to the fact that optimized path planning is the key subject in the navigating of the mobile robot. There are two different type of path planning algorithms, global and local. Whereas the global path planning requires complete information of the obstacles and environment, the local path planning needs to collect information through sensors as it moves around in unknown environments. Even though many researchers have developed various algorithms for the path planning typically including D*, Bug, Vector Field Histogram (VFH) and Dynamic Window Approach algorithm for static and dynamic unknown environment, still several particular issues, especially in construction sites, should be considered in the navigation of mobile robots based on the purpose and function of the mobile robot itself.

One of the widely used algorithms for mobile robot and autonomous vehicle navigation is D*, which has had many variations over the years. It draws a map into discrete areas called cells to finds the shortest path from its origin to destination with the assumption [5] that it does not have any obstacles in the unknown area. Each cell has a backpointer, representing the optimal traveling direction in the cell's area and costs for traveling to neighboring cells [6]. When it enters in an unknown environment or encounters obstacles, it adds new information to its map, and find a new shortest path from current location to the goal if necessary. Due to the limitations given by the cells, this has created a problem on the functionality of the D* in the setting of the map causing an out-of-bounds area. When a target is located on the out-of-bounds area, D* receives signals that the target is unreachable because there are no cells covering the area that is located on the outer part of the wall [6].

Another simple type of path planning algorithm is the Bug algorithm, where the robot moves in a straight line toward goal till it senses an obstacle. Then the robot avoids obstacles by deviating from the line while updating new important information such as new distance from the current position to the goal. Due to the simplicity of the algorithm, it requires less memory and is very competitive on total computation time compared to other algorithms. The most commonly used and referred in mobile robot path planning are Bug1 and Bug2 [7], which has weaknesses in that both algorithms do not find the optimal and shortest path in general and is not a good solution in complex areas due to its simplicity, especially in a construction site having many unpredictable objects in the environment.

Vector Field Histogram (VFH) was developed to avoid unexpected objects and easily enter narrow passages while simultaneously steering a mobile robot toward the target. A VFH-controlled mobile robot maneuvers quickly and without stopping among densely cluttered obstacles [8]. The VFH method uses a two-dimensional histogram grid for statistical representation of obstacles. The one-dimensional Polar Histogram is derived and reduced from the Histogram Grid which is updated continuously and represents the polar obstacle density around the robot in real-time. Consecutive sectors with low polar obstacle density are selected based on its near target direction. Using VFH method would not be efficient especially in construction sites requiring mapping large areas because the occupancy grid map needs a large amount of memory to be able to keep count of the values in each cell. In addition, a local minima problem, which traps the robot in another position far away from its destination causing it to travel a longer distance to reach its goal, sometimes cannot be avoided, and the method does not perform well if a narrow area has to be passed through [9].

The main idea of Dynamic Window Approach method is creating a valid search space and selecting an optimal solution from the search space. The search space is restricted to safe circular trajectories that can be reached within a short-time interval and are free from collisions [10]. The optimization goal is trying to minimize the travel time by moving fast as much as possible in the right direction. Regarding speed and safety of the mobile robot navigation in a construction site, reaching the right position efficiently to scan scenes and build a 3D environment map without any collisions is the highest priority. In addition, the local minima problem exists because of over-speed and may prevent the robot from real-time stability and non-optimal motion decision for obstacle avoidance due to not considering the size constraint of a mobile robot [11].

2.2 Object Recognition

Once 3D data of the construction environment has been collected, it is necessary to identify and localize objects in the scene that are relevant to the construction management team. Object recognition from point clouds of a construction site setting has been explored for multiple applications such as safety monitoring [12], as-built modeling [13, 14], surveying tasks [15], and performance monitoring [16]. Live construction sites are especially challenging for the object recognition task due to the unstructured nature of the site which contains many moving objects as well as articulated objects [17].

There have been several attempts to handle object recognition in construction settings from the literature. Dimitrov et al. [18] proposed a region-growing method for context-free segmentation of point cloud data. The method performed well on challenging point clouds of the built environment due to the ability to handle different sources of point cloud data variance. However, the method only extracted object contours from a scene and did not explicitly identify the types of objects. Wang et al. [19] used boundary detection methods to automatically extract Building Information Modelling (BIM) components from point cloud data. A region-growing plane segmentation step followed by a rule-based classification method was applied to identify walls, doors, windows, and roof elements from point cloud data. However, the method does not work well with incomplete data and can only be applied as a post-processing step after the whole site is scanned. Bosché et al. [20] studied a Hough transform-based method to detect Mechanical, Electrical, and Plumbing (MEP) components from laser-scan data. The method was able to account for objects that are not built in the planned location, as well as objects that may be incomplete. The work focused on the case of cylindrical MEP components and is difficult to generalize to the more general case of non-parametric objects. Chen et al. [21] proposed the use of a voxel grid-based 3D descriptor to classify different construction equipment in construction sites. A machine learning framework was used in conjunction with hand-crafted feature vectors to train a classifier from construction equipment Computer Aided Design (CAD) models. However, the work only considered outdoor environments with less clutter and only detected limited categories of construction equipment. Pu et al. [22] investigated object recognition on laser-scanned point clouds for road safety inspection. The classification algorithm uses geometric features such as size, shape, orientation and topological relationships of the point cloud segments assign detailed classes such as traffic signs, trees, building walls, and barriers. This work also only considered outdoor environments such as a roadway which has notably less obstruction than a construction site. Kim et al. [23] proposed the thermal-based object recognition. This paper introduces an innovative method for generating thermal-mapped point clouds of a robot's work environment and performing automatic object recognition with the aid of thermal data fused to 3D point clouds. However, it was limited to heat source objects, indoor environments, and it depends on the light condition.

Overall, there exist many mobile robot navigation and path planning algorithms. However, the construction sites are outdoor complex dynamic environments. The state-of-the-art path planning algorithm can occur the problem of course of dimensionality, which arises when analyzing data in high-dimensional (often with thousands) spaces. Furthermore, the state-of-the-art object detection methods are limited to post data processing and are not suitable as a source of semantic feedback during the scanning process. Thus, this study proposes the use of object detection while the laser scan is carried out to achieve real-time scene understanding for the mobile robot. Based on that information, the local planner finds a local goal and optimal path within the scanned area.

3 Methodology

3.1 Mobile Robot 3D Data Collection and Mapping

A mobile robot with a hybrid LiDAR system, GRoMI, was used for this study. GRoMI is composed of two major parts, a hybrid laser scanning system and an autonomous mobile robot platform as shown in Fig. 1. The upper part is a hybrid laser scanning system, and the lower part is a mobile robot platform. The advantages of this robotic system are: (1) data acquisition is possible while the robot is moving; (2) robot navigation and data collection can be carried out remotely or autonomously. Figure 2 describes the flowchart of the proposed approach.

First, GRoMI detects objects around it when it starts scanning. Second, the local goal position is determined by the nearest frontier algorithm [24] by using horizontally mounted laser scanners within the range of the currently built 2D map area. It is to choose next local goal as the nearest unscanned location within currently built 2D occupancy map. Then, an artificial potential field is constructed to find the best trajectory to the goal position by using the position and size of detected obstacles. While GRoMI is moving to the goal position, it calculates the current position and orientation from a currently estimated map and a previously built map of the environment with LiDAR scanners. If GRoMI reaches the local goal, it computes a new local goal position and moves toward it. This process is repeated until it reaches the global goal position.

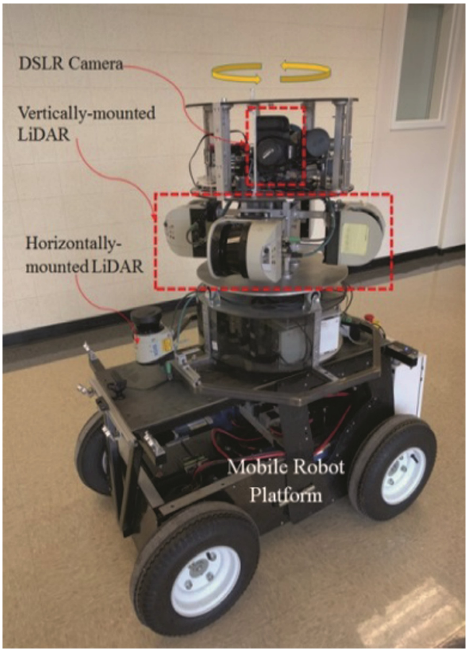


Fig. 1. Mobile robot platform with a hybrid 3D scanning system

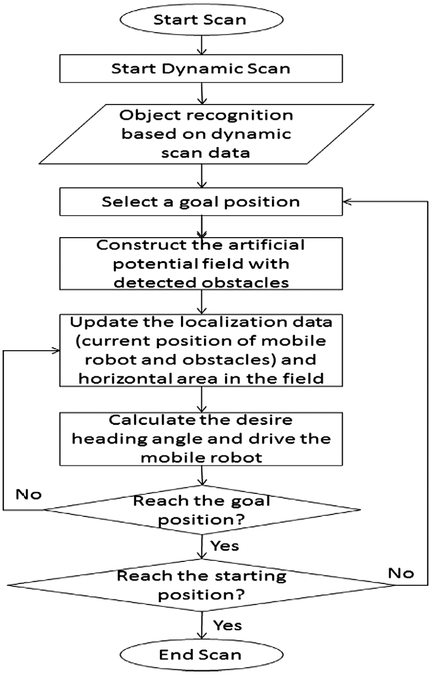
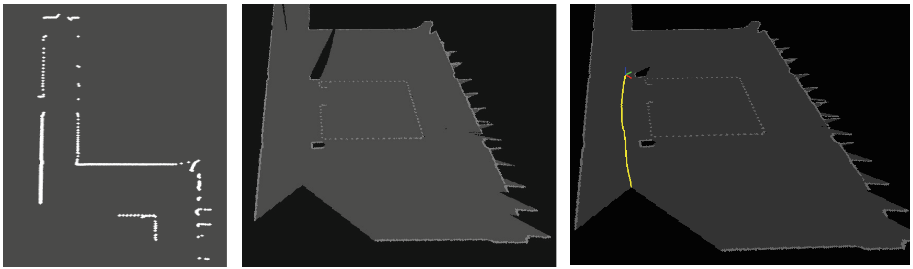

Start Scan
Start Dynamic Scan
Object recognition based on dynamic scan data
Select a goal position
Construct the artificial potential field with detected obstacles
Update the localization data (current position of mobile robot and obstacles) and horizontal area in the field
Calculate the desire heading angle and drive the mobile robot
Reach the goal position?
No
Yes
Reach the starting position?
No
Yes
End Scan

Fig. 2. Flowchart of the proposed framework

The laser scan data from the horizontal LiDAR is used by the SLAM algorithm to calculate the position and orientation of the mobile robot on the horizontal plane. The Hector SLAM algorithm, developed by Kohlbrecher et al. [25], was adopted in this study to perform laser scan matching between a current LiDAR scan and an incrementally-built map to obtain the pose estimation and horizontal map of the surrounding environment. The SLAM algorithm is performed as shown in Fig. 3. Figure 3a shows the current raw scan data from the horizontal LiDAR. Then, scan matching is performed with a currently estimated map and a previously built map of the environment in order to obtain translation and rotation parameters, which is shown in Fig. 3b. This process is continuously repeated to compute both position and orientation of the robot at each step along its trajectory as well as the 2D map of the environment, which is shown in Fig. 3c.

The position and orientation information, obtained from the 2D Hector SLAM algorithm, can be utilized to generate a 3D map by adding vertical scanning information. The vertical scans are first registered in the local coordinate frame of the mobile robot based on the angular displacement of the revolving frame of the vertical LiDARs. Then, each scan is added to the global coordinate frame by the corresponding transformation and rotation parameters estimated by the Hector SLAM. Therefore, 3D maps are gradually built up by stacking multiple planar scans.

During the SLAM process, GRoMI gets a position update every 1/25 of a second (i.e., 25 Hz) with the horizontally mounted LiDAR. The received information contains



(a) raw scan data from LiDAR

(b) incrementally-built 2D map

(c) trajectory with position and orientation of the mobile robot

Fig. 3. Scan matching with horizontal LiDAR

positions for the robot, positions of all objects on the field and the goal position. The new goal position is updated continuously when it reaches a goal position. On the other hand, GRoMI also collects data from the vertically mounted laser scanner to avoid obstacles which have distance information at different vertical angles from the mobile robot to obstacles in a 3D environment. Figure 4 illustrates this process. The numerous vertical laser lines (example lines of 30, 15, 10° are shown) are emitted from the vertical LiDARs. If there are no obstacles near the mobile robot, each line will generate perfect circles with a certain radius. On the other hand, the circles will be distorted if there exist obstacles in the environment as shown in the left figure. By using this information, the mobile robot is able to avoid obstacles.

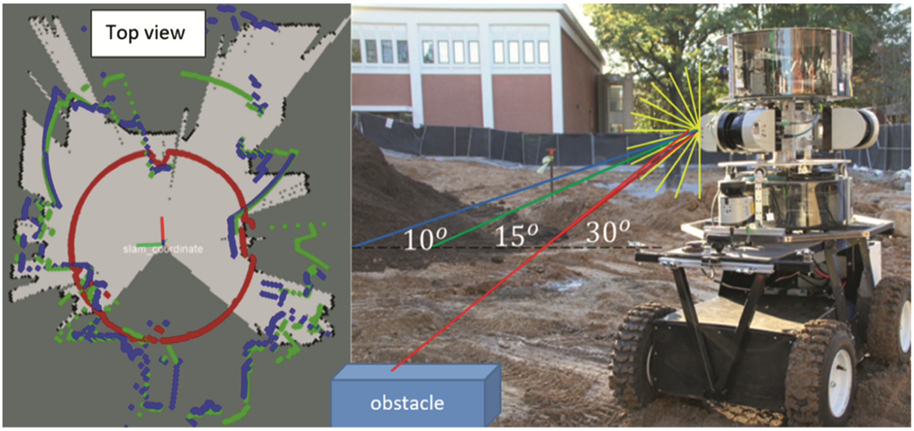


Fig. 4. Detecting obstacles by using vertical laser lines (–90° to 90°)

3.2 Object Detection

The robot navigation process is aided by an object detection routine which gives a pathfinding algorithm for a basic semantic understanding of the environment. The semantic classes assigned to the scanned environment is divided into (i) building elements such as ceiling, floor, and walls which remain static and (ii) obstacles on the ground such as equipment, materials, and workers which could potentially move around. The static elements represent absolute barriers through which the robot cannot navigate whereas the dynamic elements represent temporary barriers which have to be avoided by the robot at the current time but could disappear in the future.

A simple rule-based classifier is used since it is computationally efficient and does not require point cloud segmentation. First, a voxel-grid representation is used to down-sample the raw point cloud such that each 10 cm × 10 cm × 10 cm voxel contains a unique point. The Principal Component Analysis is used to normalize the voxel-grid coordinates such that z is zero at the elevation where the laser scanners are placed, and the x-y axes are aligned with the surrounding walls. Next, a rule-based formulation is used to classify each voxel (at position x, y, z) into 4 classes as given in the equation below. The notation I P I indicates the total number of voxels in the point cloud which satisfies a given condition. For example, voxels at a particular elevation, z', are classified into the ceiling region if the total number of voxels at that elevation exceeds the ceiling threshold. Note that this formulation allows the floor and ceiling regions to have multiple elevations since the robot may travel into different rooms with different ceiling and floor heights. In this study, the voxel thresholds are set to 50, 50, and 150 for ceiling, floor, and wall respectively

$$C_{(x,y,z)} = \begin{cases} \mathit{ceiling}, & z \in \left\{z', \left|P_{z>0\,\&\,z=z'}\right| > \mathit{threshold}_{\mathit{ceiling}}\right\} \\ \mathit{floor}, & z \in \left\{z', \left|P_{z<0\,\&\,z=z'}\right| > \mathit{threshold}_{\mathit{floor}}\right\} \\ \mathit{wall}, & x \in \left\{x', \left|P_{x=x'}\right| > \mathit{threshold}_{\mathit{wall}}\right\} \mathit{or} \\ & y \in \left\{y', \left|P_{y=y'}\right| > \mathit{threshold}_{\mathit{wall}}\right\} \\ \mathit{obstacle}, & \mathit{otherwise} \end{cases}$$

A separate obstacle detection routine is simultaneously run to handle voxels that are detected as such. The obstacle detection routine is updated every second to take into account moving objects. Neighboring voxels are iteratively merged together to form clusters of objects. Each cluster is assigned a bounding box determined by the combined dimensions of its constituent voxels. Further, the Principal Component Analysis is used to obtain the orientation of the bounding box.

3.3 Obstacle Avoidance for Navigation

From the results of obstacle detection, an obstacle avoidance algorithm can be used to safely navigate the mobile robot. The artificial potential field method proposed by Khatib [26] uses the approach of creating virtual forces. The mobile robot, obstacles, and a goal position are three basic components of this method. The basic idea is that obstacles generate an artificial repulsive force and a goal position produces an artificial attractive force. Therefore, the goal position and obstacles will generate an optimization problem which is to find a feasible trajectory from the current mobile robot position to a goal position. The direction of the mobile robot in any position of the field is decided by the direction of total field intensity, which is the summation of the artificial forces.

Figure 5 shows an example of an artificial potential field with three obstacles and the resulting trajectory from a starting position to a goal position. The starting position is (0,0) which is located in the left part of Fig. 5, and the goal position is located in the right part of Fig. 5. As shown in Fig. 5, the potential value of starting position is the highest except obstacles. On the other hand, the potential value of the goal position is the lowest. Therefore, the mobile robot flows down naturally to the goal position while avoiding obstacles because every position in the field has a gradient which points the best direction to the goal position.

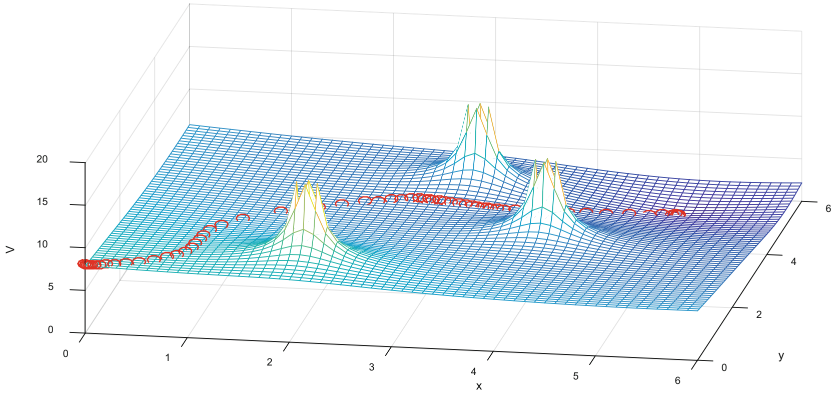


Fig. 5. Example of an artificial potential field and the resulting trajectory

The artificial attractive potential field (V_g) to a goal position from the current position of the mobile robot is defined as

$$V_g = K_g r_g \tag{1}$$

where K_g is a design parameter and r_g is defined as

$$r_g = (x_g - x)^2 + (y_g - y)^2 \tag{2}$$

which is the square of the distance between the mobile robot and the goal position. (x_g, y_g) represents the center of goal position and (x,y) is the current position of the mobile robot. The minimum of the function is reached to zero when the mobile robot approaches to a goal position.

The repulsive potential field (V_o) for an obstacle is defined as

$$V_o = \frac{K_o}{r_o} \tag{3}$$

where K_o is a design parameter and r_o is defined as

$$\mathrm{r_o} = \left(x_o - x\right)^2 + \left(y_o - y\right)^2 - R^2 \tag{4}$$

which is the square of the distance between the mobile robot and the obstacle position minus a square of radius R. The radius R is room to avoid the collision which is set as 0.5 m. (x_o,y_o) represents the center of obstacle position and (x,y) is the current position of the mobile robot. In this method, each basic component is modeled as a point. However, the areas where the mobile robot should not be reached are needed to be represented as a boundary. Therefore, the radius R is included in the modeling equation, which information is given by the object detection described in the following section. Also, the Eq. (4) generates a pillar barrier with infinite height to move the mobile robot without crossing any of the boundaries. In this paper, the parameter $\mathrm{K_g}$ is fixed to 1 and $\mathrm{K_o}$ is changed while the mobile robot is moving. The parameter $\mathrm{K_o}$ determines the importance of obstacles, which is calculated by the distance between the mobile robot and the obstacles. The higher weighting obstacles, those that are close to the mobile robot, are taken into account as important obstacles when calculating potential field. The total potential field of the local scene is computed by adding individual potential field values as Eq. (5).

$$\mathrm{V} = \mathrm{V_g} + \sum_{i=1}^{n} V_{o_i} \tag{5}$$

The distance the mobile robot moves toward or away from the obstacles is used to increase or decrease the parameter $\mathrm{K_o}$ and thus making the change of total potential field values.

$$\begin{aligned} \mathrm{H_o} &= \mathrm{K_o} + e^{-a \cdot d}, d = \sqrt{\left(x - x_{o_i}\right)^2 + \left(y - y_{o_i}\right)^2 < 0} \\ \mathrm{K_o} &= \mathrm{K_o} - e^{-a \cdot d}, d = \sqrt{\left(x - x_{o_i}\right)^2 + \left(y - y_{o_i}\right)^2 > 0} \end{aligned} \tag{6}$$

where, a is a factor which is set to 2. The best range for the parameter $\mathrm{K_o}$ is between 0.05 and 1 because the parameter should be positive and not be too large. Therefore, the parameter should be set to 0.05 when it becomes smaller than 0.05 and set to 1 when it becomes larger than 1.

The trajectory should be determined by the best direction to the goal position for the current position of the mobile robot. The Newton direction method is used in this paper, which represents the optimal direction in which a step should be taken next. The Newton direction is defined as

$$\vec{s} = -H_V^{-1} \cdot \nabla V \tag{7}$$

where ∇V is the gradient which describes the first order derivative of the total potential field and H_V^{-1} is the inverse of the Hessian matrix which describes the second order derivative of the total potential field in Eq. (5). The first order derivative determines the steepest descent but it is not necessarily the goal direction. Therefore, the second derivative information is important due to finding optimal trajectory when the slope is tilting left or right. This is because the slope is not real but an artificial potential field.

4 Simulation and Experimental Results

Simulated data was first used to evaluate the artificial potential field method. As shown in Fig. 6, the laser scanning process was simulated by the yellow line which detects obstacles. In this simulation, the moving obstacles were used to demonstrate the robustness of this method. The black circle is the original position of obstacles, and the red circle is the current position of moving obstacles as times goes by. The light blue circle, near the red circle which is the current position of the mobile robot, represents the estimated position of obstacles sensed by LiDAR. Lastly, the light red circle indicates the

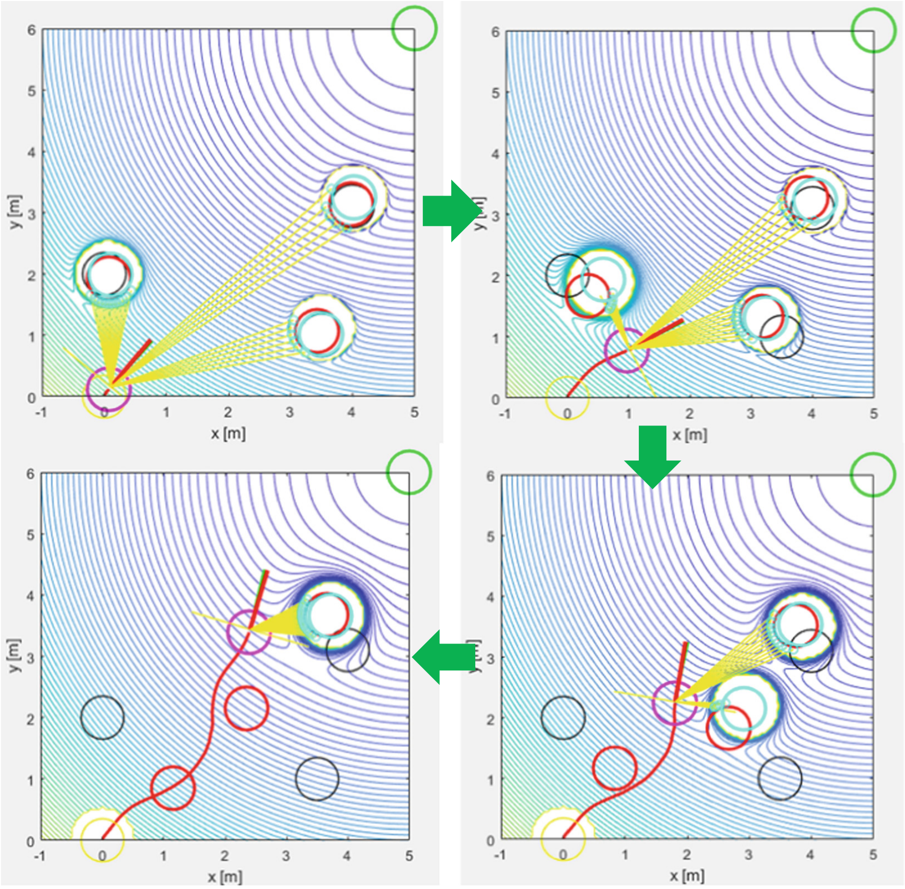


Fig. 6. Simulation combining the artificial potential field and odometry model (Color figure online)

current position of the mobile robot and the red line represents the optimal trajectory of mobile robot calculated by the artificial potential field. The contour line demonstrates the level of potential in every point in the field which has the pillar barrier obstacles and the local minimum at a goal position.

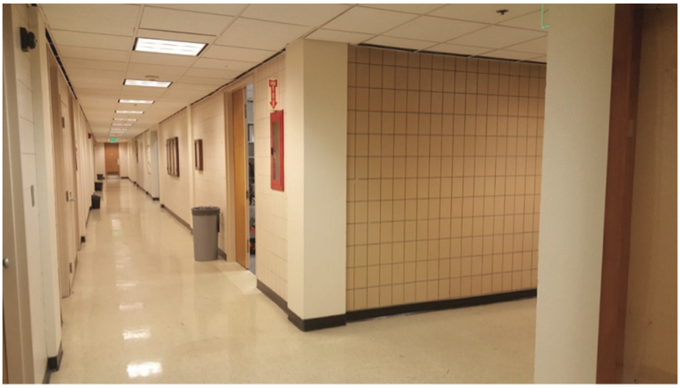
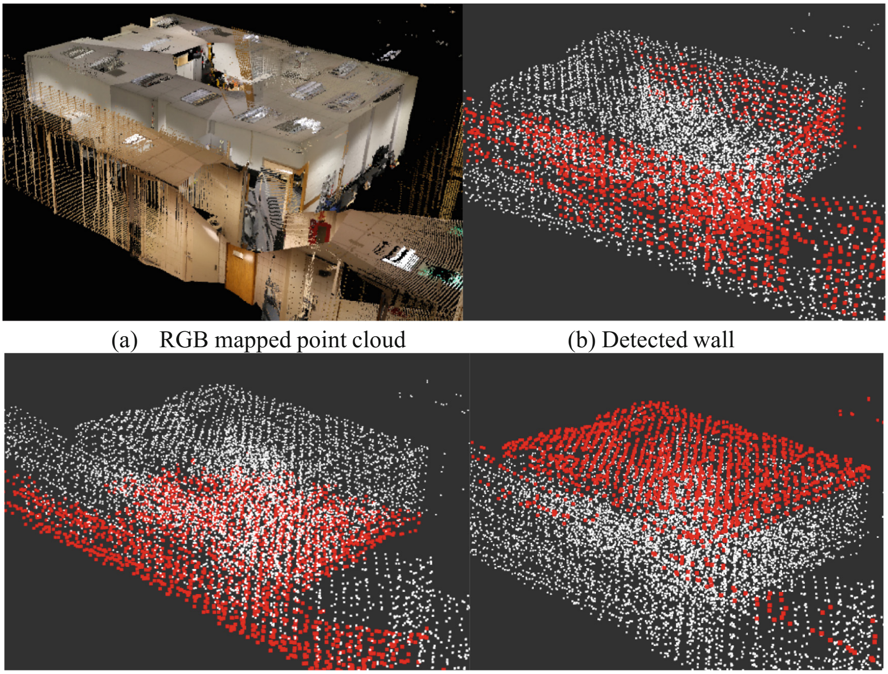


Fig. 7. Image of the hallway where indoor laser scan experiment was carried out



(a) RGB mapped point cloud

(b) Detected wall

(c) Detected floor

(d) Detected ceiling

Fig. 8. Building element recognition

Based on this simulation result, the navigation and obstacle avoidance algorithm with artificial potential field method was implemented to GRoMI. The object detection and the navigation routine were evaluated through a series of laser scanning experiments conducted in an indoor environment. The building consists of a series of hallways, an example of which is shown in Fig. 7. Figure 8 shows the recognition result of building elements. The scan results are continuously updated in real time as the robot progresses through the environment.

Figure 9 shows the obstacle detection results, where the red boxes indicate individual obstacle instances that have to be avoided by the mobile robot, while the white points indicate the raw point cloud data, and the bold yellow line indicates the trajectory of the robot. The obstacle position and orientation are updated at every scan cycle to take into account possible movement of objects. The updated position and volume data of obstacles are provided to the navigation routine for constructing the obstacle barrier on the potential field. Only coarse object bounding boxes for detected obstacles are needed for navigation purposes.

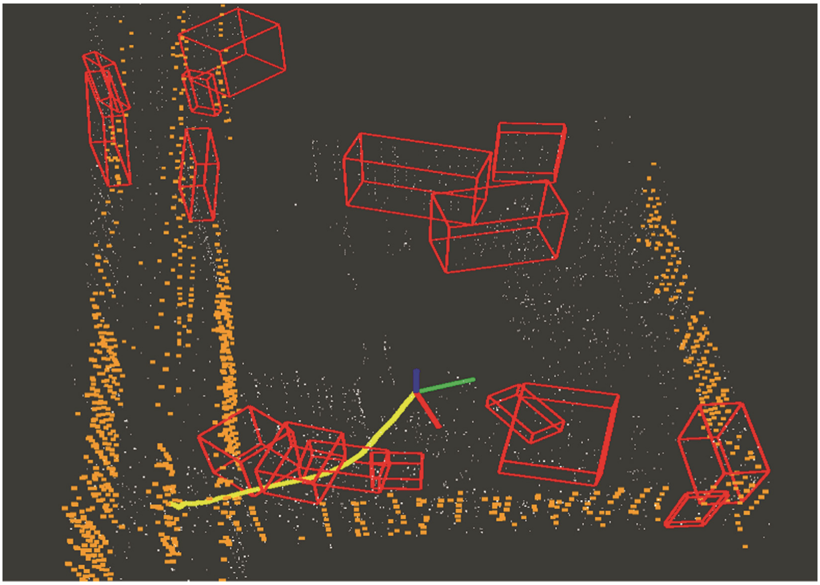


Fig. 9. Obstacle detection in the indoor environment (Color figure online)

With the same manner, an outdoor construction site was tested. Figure 10 shows the final 3D point clouds with fused RGB data which were generated by GRoMI. The bold yellow line indicates the trajectory path that GRoMI created. The 3D point clouds were built up by capturing laser scanner and camera data in static scan positions. Four separate scans were combined into a single point cloud automatically using the transformation matrices derived from the SLAM solution.

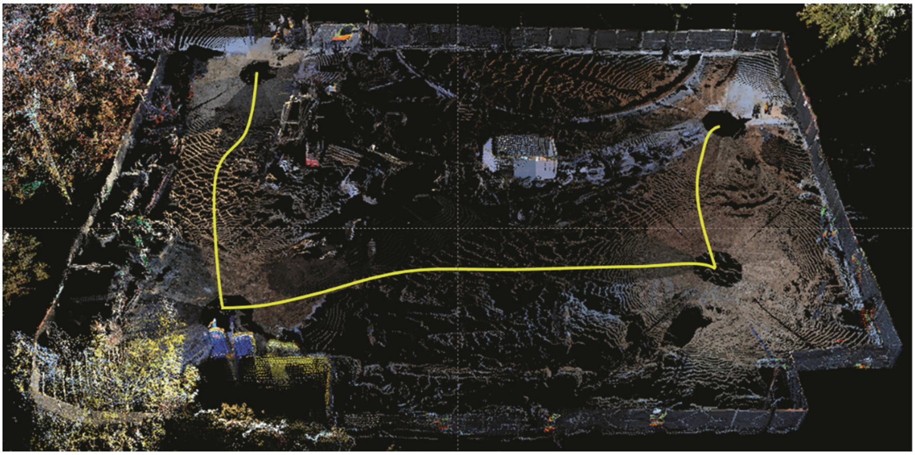


Fig. 10. An autonomously generated 3D point cloud of the unstructured outdoor construction site by GRoMI. A bold yellow line is the trajectory path created by GRoMI. (Color figure online)

5 Conclusion and Discussion

This study proposed and implemented a custom-designed mobile robot platform for use in automated 3D data collection from construction sites. The mobile robot used Simultaneous Localization and Mapping (SLAM) coupled with object detection and obstacle avoidance algorithms to successfully navigate unstructured and cluttered environments. The SLAM result was used to localize the mobile robot in the scanning environment and determine the translation and rotation of each set of scans in the world coordinate system. Whereas, the object detection results were used to rapidly recognize static obstacles such as walls and dynamic obstacles such as equipment, materials, and workers. At the end of the automated scanning process, a complete and registered 3D point cloud of the scanning environment was obtained.

Experimental results showed that the proposed approach was an effective way of detecting and avoiding obstacles in real-time for a mobile robot to successfully collect as-is 3D geometry data at both indoor and outdoor construction sites. Also, it was identified that the outdoor application was much more challenging for the autonomous mobile robot compared to the indoor application due to uneven ground conditions, different soil types and conditions (e.g., loose, wet), and many scattered obstacles on the ground. Especially, dead-reckoning could not be applied at the outdoor site due to loose and slippery soil conditions, which is very common in construction sites. The 2D SLAM and 3D laser scanning for navigation, however, were sufficiently enough for estimating robot's location and navigation paths. Overall, the point clouds collected by GRoMI were sufficiently accurate to be used for many construction management applications such as construction progress monitoring, safety hazard identification, defect detection, and material inventory management. For future study, the research team plans to investigate the collaborations between laborers and an autonomous mobile robot in close proximity with potential field construction applications. Moreover, machine learning techniques will be further explored to improve the current point cloud object detection accuracy.

References

1. Cho, Y.K., Wang, C., Tang, P., Haas, C.T.: Target-focused local workspace modeling for construction automation applications. J. Comput. Civ. Eng. **26**(5), 661–670 (2012)
2. Pătrăucean, V., Armeni, I., Nahangi, M., Yeung, J., Brilakis, I., Haas, C.: State of research in automatic as-built modelling. Adv. Eng. Inform. **29**, 162–171 (2015)
3. Chen, J., Cho, Y.K.: Real-time 3D mobile mapping for the built environment. In: 33rd International Symposium on Automation and Robotics in Construction (2016)
4. Kim, P., Chen, J., Cho, Y.K.: SLAM-driven robotic mapping and registration of 3D point clouds. Autom. Constr. **89C**, 38–48 (2018)
5. Koenig, S., Tovey, C., Smirnov, Y.: Performance bounds for planning in unknown terrain. Artif. Intell. **147**(1–2), 253–279 (2003)
6. Ng, J.: A Practical Comparison of Robot Path Planning Algorithms given only Local Information (2012)
7. Lumelsky, V.J., Stepanov, A.: Dynamic path planning for a mobile automaton with limited information on the environment. IEEE Trans. Autom. Contr. **31**(11), 1058–1063 (1986)
8. Borenstein, J., Koren, Y.: Real-time obstacle avoidance for fast mobile robots in cluttered environments. In: IEEE International Conference on Robotics and Automation, pp. 572–577, May 1990
9. Sobh, T., Xiong, X.: Prototyping of Robotic Systems: Applications of Design and Implementation, p. 498. Hershey, Information Science Reference (2012)
10. Fox, D., Wolfram, B., Thrun, S.: The dynamic window approach to collision avoidance. IEEE Robot. Autom. Mag. **4**, 1–23 (1997)
11. Saranrittichai, P., Niparnan, N., Sudsang, A.: Robust local obstacle avoidance for mobile robot based on Dynamic Window approach. In: 2013 10th International Conference on Electrical Engineering, Computer, Telecommunications and Information Technology ECTI-CON 2013, no. 1, pp. 4–7 (2013)
12. Chi, S., Caldas, C.H.: Automated object identification using optical video cameras on construction sites. Comput. Civ. Infrastruct. Eng. **26**(5), 368–380 (2011)
13. Tang, P., Akinci, B., Huber, D.: Semi-automated as-built modeling of light rail system guide beams. In: Proceedings of Construction Research Congress 2010, vol. 373, no. 41109 (2010)
14. Jung, J., Hong, S., Yoon, S., Kim, J., Heo, J.: Automated 3D wireframe modeling of indoor structures from point clouds using constrained least-squares adjustment for as-built BIM. J. Comput. Civ. Eng. **30**(1), 4015074 (2015)
15. Siebert, S., Teizer, J.: Mobile 3D mapping for surveying earthwork projects using an unmanned aerial vehicle (UAV) system. Autom. Constr. **41**, 1–14 (2014)
16. Yang, J., Park, M.-W., Vela, P.A., Golparvar-Fard, M.: Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future. Adv. Eng. Inform. **29**(2), 211–224 (2015)
17. Wang, C., Cho, Y.K.: Performance test for rapid surface modeling of dynamic construction equipment from laser scanner data. In: ISARC Proceedings (2014)
18. Dimitrov, A., Golparvar-Fard, M.: Segmentation of building point cloud models including detailed architectural/structural features and MEP systems. Autom. Constr. **51**(C), 32–45 (2015)

19. Wang, C., Cho, Y.K., Kim, C.: Automatic BIM component extraction from point clouds of existing buildings for sustainability applications. Autom. Constr. **56**, 1–13 (2015)
20. Bosché, F., Ahmed, M., Turkan, Y., Haas, C.T., Haas, R.: The value of integrating Scan-to-BIM and Scan-vs-BIM techniques for construction monitoring using laser scanning and BIM: the case of cylindrical MEP components. Autom. Constr. **49**, 201–213 (2015)
21. Chen, J., Fang, Y., Cho, Y.K., Kim, C.: Principal axes descriptor for automated construction-equipment classification from point clouds. J. Comput. Civ. Eng. **31**, 1–12 (2016)
22. Pu, S., Rutzinger, M., Vosselman, G., Oude Elberink, S.: Recognizing basic structures from mobile laser scanning data for road inventory studies. ISPRS J. Photogramm. Remote Sens. **66**(6 SUPPL), S28–S39 (2011)
23. Kim, P., Chen, J., Cho, Y.K.: Robotic sensing and object recognition from thermal-mapped point clouds. Int. J. Intell. Robot. Appl. **1**(3), 243–254 (2017)
24. Freda, L., Oriolo, G.: Frontier-based probabilistic strategies for sensor-based exploration. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 3881–3887 (2005)
25. Kohlbrecher, S., Von Stryk, O., Meyer, J., Klingauf, U.: A flexible and scalable SLAM system with full 3D motion estimation. In: Proceedings of the 2011 IEEE International Symposium Safety, Security, and Rescue Robotics, Kyoto, Japan, 1–5 November, pp. 155–160 (2011)
26. Khatib, O.: Real-time obstacle avoidance for manipulators and mobile robots. Int. J. Robot. Res. **5**(1), 90–98 (1986)

Computer Supported Construction Management

Making Each Workhour Count: Improving the Prediction of Construction Durations and Resource Allocations

Martin Fischer[1](✉), Nelly P. Garcia-Lopez[2], and René Morkos[3]

[1] Department of Civil and Environmental Engineering, Stanford University,
473 via Ortega, Room 297, Stanford, CA 94305-4020, USA
fischer@stanford.edu

[2] Grupo Galopa, Cl. 119 #1143, Bogotá, Colombia
nellygarci@gmail.com

[3] ALICE Technologies, 1040 Noel Dr Suite 203, Menlo Park, CA 94025, USA
rene@alicetechnologies.com

Abstract. Construction duration is an important performance aspect of building projects because it determines the time-to-market for a building, i.e., it stands between a client's final decision to construct a building and obtaining the benefits from the designed building. This paper shows how intelligent computing (including semantic modeling, simulation, genetic algorithms, and machine learning) improves the ability of construction professionals to predict the construction schedule duration and direct cost of building projects at the beginning of construction and during construction. Such predictions are important because they inform the schedule commitments made and the allocation of resources that practitioners believe will let them meet the commitments. Hence, the prediction methods must capture the most important phenomena that are likely to impact the duration of activities and of construction. The two applications discussed – Tri-Constraint Method (TCM) and Activity-Flow Model (AFM) – incorporate key phenomena observed in practice, such as handling of workspace constraints and flows required for the execution of activities, that are not part of the currently prevalent concepts and tools. TCM and AFM significantly improve the ability of construction professionals to make more reliable predictions of construction duration.

1 Introduction

The direct cost of a construction project is determined by the sum of the efforts required to execute all construction activities and the duration is determined by the sum of all the activity durations minus the overlaps of activities. Cost and duration depend chiefly on (1) the design of the structure to be built, (2) the selection of construction methods, (3) the sequence of the activities, and (4) the conditions (e.g., information and material availability, access to the workface) for the execution of the construction activities. These four topics interact. For example, a particular sequence of activities leads to certain conditions on site, which affect the effort required for the completion of the activities and therefore the cost and duration of the activities and consequently the

© Springer International Publishing AG, part of Springer Nature 2018

I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 273–295, 2018.
https://doi.org/10.1007/978-3-319-91635-4_15

project. Current tools to schedule and manage construction activities fall short in their support of construction professionals to create and adjust a short and cost-effective construction schedule in a timely manner.

For example, for a construction activity to be executable, it must meet several types of constraints. It must be placed in the right sequence of work, i.e., it must respect the work logic. These constraints are handled well by the Critical Path Method (CPM), which is a common method to predict the duration of construction (Fondahl 1961). The resources necessary for an activity's execution must also be available. Of particular importance are the work crews and equipment that transform the input materials into the desired output and the corresponding workspaces needed. The crew constraint is challenging because there is typically some flexibility in terms of the number of crews available. The workspace constraint is challenging because it is a unary resource, i.e., at a particular workface, there is only one workspace available, which means that it needs to be formalized as a disjunctive constraint. These constraints interact in complex ways. The number of crews determines the number of workspaces needed. The workspaces available determines the number of crews that can work safely and productively. The number of crews and workspaces determines feasible sequences of construction, which, in turn determine the duration and cost of construction. The Location-Based Scheduling (LBS) method only partially addresses these constraints by breaking up a building project into different work locations (Kenley and Seppänen 2009).

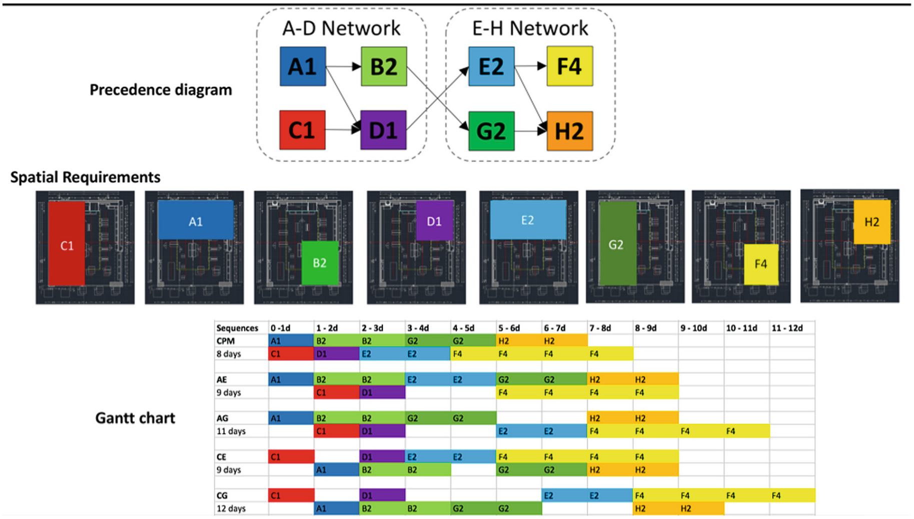
Furthermore, in addition to the resources and workspaces, activities need materials, labor, information, permits, etc. to arrive in a timely manner. The site managers don't only need to manage the activities, but also the flows that enable the activities. Koskela (1999) conceptualized these flows into seven categories (labor, equipment, workspace, materials, precedence, information, and external flows), but they had not been formalized so that they can be implemented in a tool. In addition, activity durations and flow arrival dates and times are uncertain, i.e., their variability must be understood and managed. Current methods, e.g., the Last Planner System (LPS) (Ballard and Howell 1998), enable the partial understanding of this variability, but don't offer predictions for expected variability and the likely causes of variability. As a consequence, one observes large time buffers, which extend the project duration, or costly inventory (work-in-progress (WIP)), and occasionally capacity buffers, on site.

This paper describes the application of intelligent computing to improve the sequencing of construction activities for a given design and set of construction methods and increase the understanding of variability of flows and activities to improve the initial prediction of construction duration and cost and enhance the ability of site managers to rapidly and intelligently adjust construction schedules based on the current situation on site. These applications were tested on six ongoing construction projects.

2 Parametric Construction Scheduling

2.1 Types of Constraints for Construction Scheduling

A good construction schedule includes all the right activities in the right sequence. Ideally, construction schedules would be generated automatically based on key input parameters and general and project-specific constraints. The right activities are determined by the scope to be built, typically represented with a BIM (Building Information Model), and the construction methods selected. The right sequence depends on the required sequence logic between the activities and the resources (typically crews) and workspaces available. Our observations of scheduling practices have shown that only one or just a few schedules are created on construction projects. However, there are many possible sequences, which remain largely unexplored. A simple example illustrates this point (Fig. 1).



Sequences	0 -1d	1 - 2d	2 - 3d	3 - 4d	4 - 5d	5 - 6d	6 - 7d	7 - 8d	8 - 9d	9 - 10d	10 - 11d	11 - 12d
CPM	A1	B2	B2	G2	G2	H2	H2					
8 days	C1	D1	E2	E2	F4	F4	F4	F4				
AE	A1	B2	B2	E2	E2	G2	G2	H2	H2			
9 days		C1	D1			F4	F4	F4	F4			
AG	A1	B2	B2	G2	G2			H2	H2			
11 days		C1	D1			E2	E2	F4	F4	F4	F4	
CE	C1		D1	E2	E2	F4	F4	F4	F4			
9 days		A1	B2	B2		G2	G2	H2	H2			
CG	C1		D1				E2	E2	F4	F4	F4	F4
12 days		A1	B2	B2	G2	G2			H2	H2		

Fig. 1. Example showing four feasible sequences to a simple construction scenario.

Given the logic of work and the workspaces required and available, there are several possible sequences with different durations. With the practice of creating one schedule, the predicted duration of this small project depends on the selection of the starting sequence by the scheduler, starting with A1 first or with C1. Note that the CPM schedule, which ignores the workspace constraints, is infeasible since crews would be occupying the same workspace at the same time.

This simple example underscores the importance of considering workspace as a constraint for construction schedules. Almost all construction projects we have studied have vast empty spaces punctuated with pockets of intense work. Construction site space, it seems, is under-utilized. So much empty space on construction sites suggests a significant potential to schedule more construction activities concurrently to reduce construction duration. However, simply increasing space utilization by scheduling more construction processes concurrently can lead to spatial congestion (Thomas et al. 2006), which is detrimental to productivity, safety, and quality (Riley and Sanvido 1995). Thus, for every construction project, there is a balanced level of construction site space utilization that achieves a short construction duration without spatial congestion. Attaining this ideal level of space utilization requires a systematic approach to construction space allocation. However, current construction management theory cannot achieve these balanced levels of space utilization because the constraint for resolving spatial requirements has not been sufficiently formalized for doing so. Morkos (2014) formalized the spatial constraint as a disjunctive constraint, which ensures that activities don't occur in the same space at the same time.

Morkos combined the spatial constraint with precedence and discrete resource capacity constraints into a Tri-Constraint Method (TCM). Precedence constraints ensure that the laws of physics and the logic of work are not violated while discrete resource capacity constraints ensure that discrete resources, such as labor, are not over-allocated. The mechanisms for resolving these three constraint types are combined with a mechanism for varying sequence, which enables TCM to generate multiple construction sequencing alternatives. To validate TCM, three construction project case studies were used to compare the best of 10,000 TCM schedules to schedules created with the Critical Path Method (CPM) and the Line of Balance (LOB) or Location-based Scheduling Method (LBMS) (Kenley and Seppänen 2006). See (Morkos 2014) for details. The CPM construction durations for the three case studies were on average 49% shorter than the fastest TCM schedules because CPM does not model spatial requirements leading to infeasible schedules with crews assigned to work in the same space. The fastest TCM construction durations were on average 45% shorter than the fastest LBMS construction durations because the LBMS scheduling method cannot schedule more than one process in a zone while guaranteeing no spatial congestion.

TCM builds on Waugh's (1990) scheduling algorithm. It creates multiple schedules that satisfy precedence, discrete, and the disjunctive spatial constraints. To do so, TCM starts all construction operations in the TODO list and uses an algorithm to move them one by one into the CANDO, DOING, and DONE lists, respectively. When all construction processes are scheduled and thus in the DONE list, the construction project is complete and the algorithm resets itself by placing all processes back in the TODO list. In the next run, the construction sequence is varied, so that a scheduler can explore as many schedule sequences as desired and understand their duration and cost implications.

2.2 Application of TCM

TCM has been implemented in a software tool called ALICE Technologies[1]. This has allowed us to test the application of the TCM on a number of projects. Figure 2 shows the representation of construction knowledge in the form of a recipe, in this case, how to build a slab. These recipes are akin to fragnets (Dong 2012). Such construction knowledge is a key input to the TCM-based scheduling method. The user indicates which construction recipes apply to which types of building elements. Given this knowledge and information and respecting gravity and resource availability constraints,

[1] The TCM research was performed by the third author in the research lab of the first author. ALICE Technologies was started by the third author. The first author serves as an advisor to ALICE Technologies.

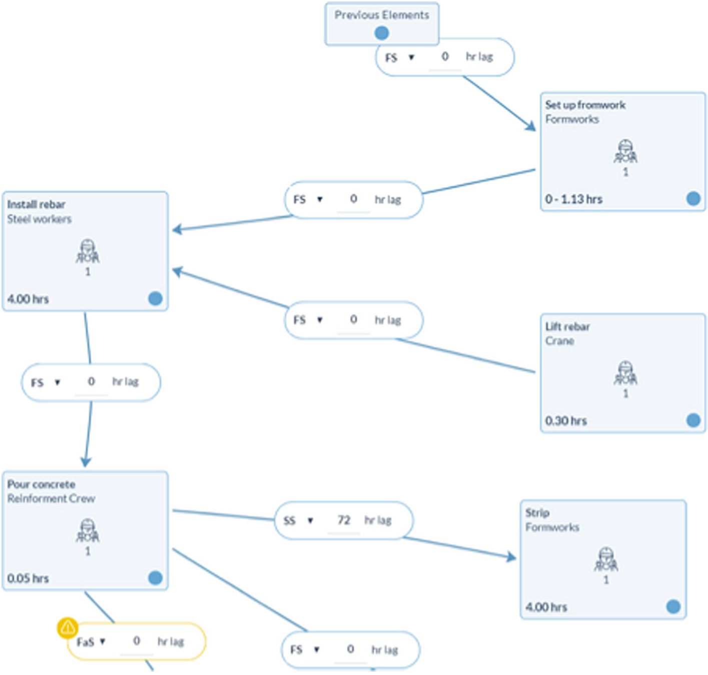


Fig. 2. Partial recipe or fragnet showing a construction method for a concrete slab.

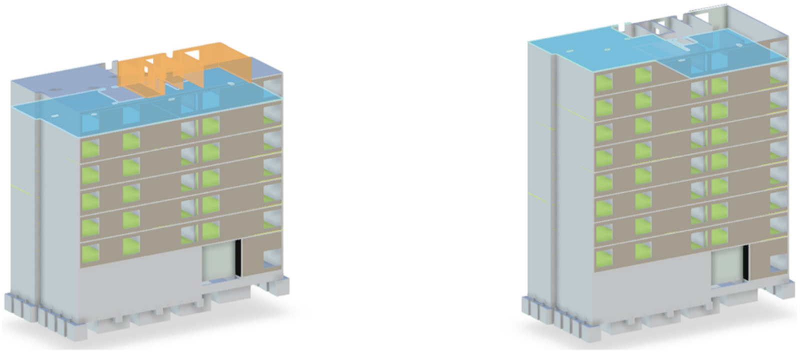


Fig. 3. 4D models generated automatically from the recipes and the project BIM.

the TCM can generate feasible construction schedules automatically. Figures 3 and 4 illustrate two key outputs produced by this scheduling method. Each schedule generated can be viewed as a 4D model (Fig. 3), since the activities from the recipes and the building elements to which the activities apply are connected. Each schedule or sets of schedule scenarios can also be compared with other schedules in a time-cost graph (Fig. 4). For example, in Fig. 4, the diamond-shaped dots show the duration and cost of five of the best baseline schedules, the square dots show schedules based on a different

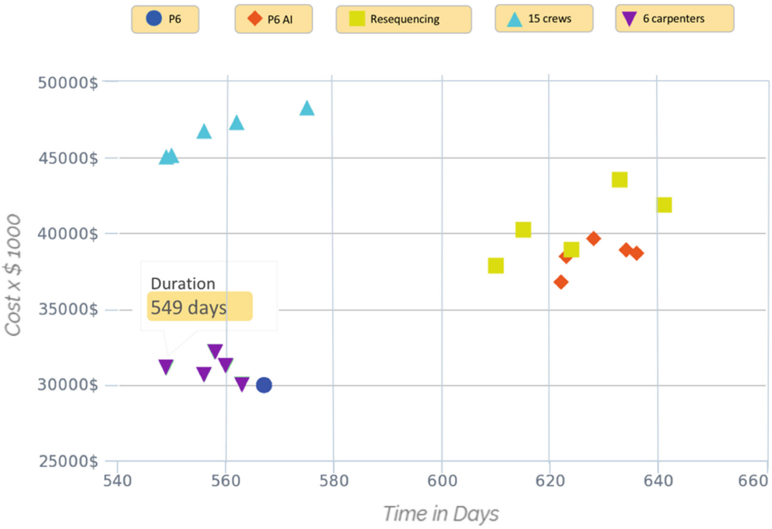


Fig. 4. Time-cost tradeoff chart produced automatically with the TCM. Each cluster of dots represents a different schedule scenario with the best five of the thousands of schedule options that are generated shown. For each dot, the corresponding 4D model, resource allocation, and other schedule information can be inspected.

sequence of façade work, the triangular dots pointing upwards show schedules for which resources were made available whenever needed, and the triangular dots pointing downwards show schedules for which the number of crews available for each trade has been optimized for duration and cost. The round dot shows the original duration and cost as predicted with the original CPM schedule, which was, however not feasible because it didn't contain all the activities and included several sequencing mistakes.

After setup of the initial parametric schedule model based on the BIM and the construction recipes, these explorations can happen over the course of just a few hours, including conversations about the pros and cons of various construction strategies.

2.3 Implications of TCM

The TCM formalizes and combines the resolution mechanisms for three constraint types. These resolution mechanisms are clearly and separately defined which could allow future construction researchers to improve these mechanisms. To find shorter feasible schedules, TCM can generate thousands of feasible schedules in minutes compared to the hours required to produce a single schedule with today's construction scheduling methods. For three case study projects, TCM produced schedules without spatial conflicts with construction durations that are 45% shorter than those produced with the LBMS method, today's best scheduling method that consider spatial requirements of construction processes to some extent. Quickly generating large numbers of feasible schedules can enable construction managers to select schedule alternatives with better schedule metric values, such as lower cost or shorter duration. This approach to construction scheduling of visualizing, manipulating, and navigating the state space acknowledges the scheduling state space which is not commonly done in industry today. This approach to construction scheduling offers a more accurate representation of the construction scheduling problem than that offered by creating and analyzing single schedules, making it more likely that a good schedule with short duration and low cost is created and maintained over the course of a project.

We consider TCM an intelligent computing method because it incorporates knowledge about construction sequencing constraints that has not been formalized before in a way to scale to full-scale building projects and it implements a simulation-based approach to identifying schedules with short duration, low cost, even resource use, or smooth workflow. Extensions could include further refinement of the sequencing knowledge (e.g., a generalized support constraint), the incorporation of other concerns like safety, and the optimization of schedules that considers different design versions or construction methods.

However, our research has not only shown limitations in today's scheduling methods (CPM, LBMS) to generate short and cost-effective schedules, but – as will be discussed in the next section – it has also shown the limitations of predicting which activities are most likely to be delayed on the basis of schedules that represent activities only.

3 Predicting Activities that Are Most Likely to Be Delayed

3.1 Why Flows Matter

The generation of the initial construction schedule and, therefore, the initial prediction of the duration and cost of construction also require an understanding of the variability of activities, which, in turn, depends on the reliability of the flows required by an activity and the reliability of the execution of an activity. This information is also needed when site managers plan the work for the upcoming weeks and allocate the available resources to make every work hour productive.

So that an activity can be done, it must have all the required flows available. Ensuring the availability of these flows is typically part of the make-ready process (Ballard and Howell 1995). These flows are often also used to categorize the reasons why an activity could not be done as planned. Figure 5 shows reasons for non-completions on projects using many of Koskela's seven flow categories (Koskela 1999).

Utilizing machine learning methods, we attempted to predict activities that are likely to be delayed based on a large dataset with data similar to those shown in Fig. 5 (Garcia-Lopez and Fischer 2016). It was, however, impossible to predict likely delays with any reliability. We realized that reducing flows to constraints on activities loses the information about where each flow is coming from, i.e., the network of flows that exists on projects is lost, which then makes reasoning about the flows and therefore the activity constraints extremely difficult. Garcia-Lopez (2017) formalized the Activity-Flow Model (AFM) to allow construction site managers to track not only activity progress and completion, but also the availability of flows as equipment and labor, for example, flow from one activity to the next (Fig. 6).

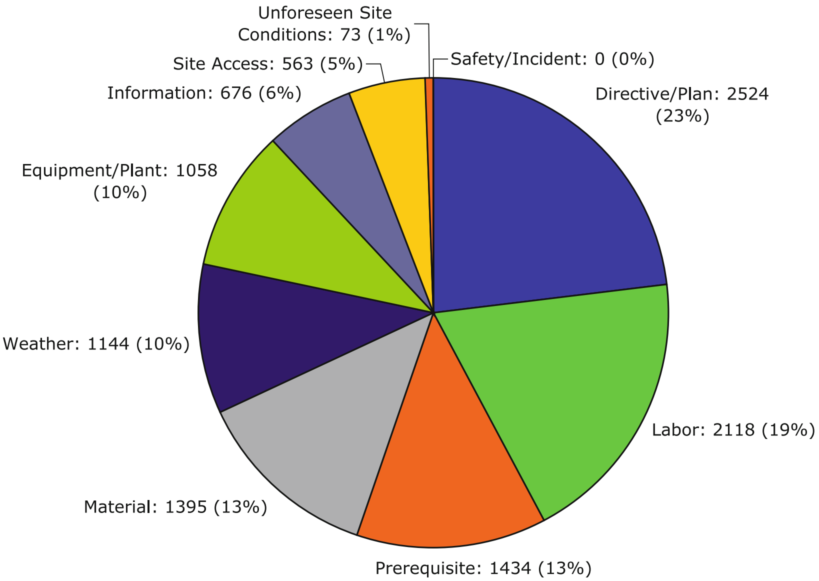
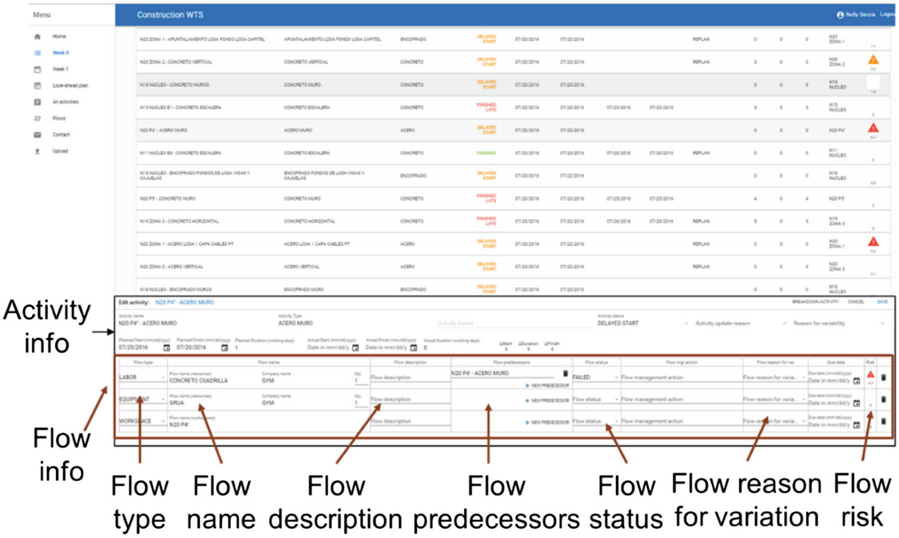


Fig. 5. Reasons for not completing an activity on time (compiled by Strategic Project Solutions, San Francisco and presented by Williams (2018)).



N20 ZONA 1 - APUNTALAMIENTO LOSA FONDO LOSA CAPITEL	APUNTALAMIENTO LOSA FONDO LOSA CAPITEL	ENCOFRADO	DELAYED START	07/20/2016	07/20/2016			REPLAN	0	0	0	N20 ZONA 1	1.1
N20 ZONA 2 - CONCRETO VERTICAL	CONCRETO VERTICAL	CONCRETO	DELAYED START	07/20/2016	07/20/2016			REPLAN	0	0	0	N20 ZONA 2	⚠ 3.2
N18 NUCLEO - CONCRETO MUROS	CONCRETO MURO	CONCRETO	DELAYED START	07/20/2016	07/20/2016				0	0	0	N18 NUCLEO	1.5
N15 NUCLEO E1 - CONCRETO ESCALERA	CONCRETO ESCALERA	CONCRETO	FINISHED LATE	07/20/2016	07/20/2016	07/23/2016	07/23/2016		3	0	3	N15 NUCLEO	0
N20 P4' - ACERO MURO	ACERO MURO	ACERO	DELAYED START	07/20/2016	07/20/2016				0	0	0	N20 P4'	⚠ 6.7
N11 NUCLEO E6 - CONCRETO ESCALERA	CONCRETO ESCALERA	CONCRETO	FINISHED	07/20/2016	07/20/2016	07/20/2016	07/20/2016	REPLAN	0	0	0	N11 NUCLEO	0
N18 NUCLEO - ENCOFRADO FONDOS DE LOSA VIGAS Y CAJUELAS	ENCOFRADO FONDOS DE LOSA VIGAS Y CAJUELAS	ENCOFRADO	DELAYED START	07/20/2016	07/22/2016				0	0	0	N18 NUCLEO	5.3
N20 P5' - CONCRETO MURO	CONCRETO MURO	CONCRETO	FINISHED LATE	07/20/2016	07/20/2016	07/25/2016	07/25/2016		4	0	4	N20 P5'	0
N19 ZONA 3 - CONCRETO HORIZONTAL	CONCRETO HORIZONTAL	CONCRETO	FINISHED LATE	07/20/2016	07/20/2016	07/26/2016	07/26/2016	REPLAN	5	0	5	N19 ZONA 3	0
N20 ZONA 1 - ACERO LOSA 1 CAPA CABLES PT	ACERO LOSA 1 CAPA CABLES PT	ACERO	DELAYED START	07/20/2016	07/20/2016			REPLAN	0	0	0	N20 ZONA 1	⚠ 5.4
N20 ZONA 3 - ACERO VERTICAL	ACERO VERTICAL	ACERO	DELAYED START	07/20/2016	07/20/2016			REPLAN	0	0	0	N20 ZONA 3	3.1
N18 NUCLEO - ENCOFRADO MUROS	ENCOFRADO MURO	ENCOFRADO	DELAYED START	07/20/2016	07/20/2016				0	0	0	N18 NUCLEO	

Edit activity: N20 P4' - ACERO MURO   BREAKDOWN ACTIVITY CANCEL SAVE

Activity name: N20 P4' - ACERO MURO
Activity Type: ACERO MURO
Activity Parent
Activity status: DELAYED START
Activity update reason
Reason for variability

Planned Start (mm/dd/yyyy): 07/20/2016
Planned Finish (mm/dd/yyyy): 07/20/2016
Planned Duration (working days): 1
Actual Start (mm/dd/yyyy): Date in mm/dd/y
Actual Finish (mm/dd/yyyy): Date in mm/dd/y
Actual Duration (working days): 0
ΔStart 0 ΔDuration 0 ΔFinish 0

Flow type	Flow name			Flow description	Flow predecessors	Flow status	Flow mgt action	Flow reason for var	Due date	Risk
LABOR	Flow name (resources): CONCRETO CUADRILLA	Company name: GYM	Qty: 1	Flow description	N20 P4' - ACERO MURO + NEW PREDECESSOR	FAILED	Flow management action	Flow reason for varia...	Due date (mm/dd/yyyy): Date in mm/dd/y	⚠ 4.3
EQUIPMENT	Flow name (resources): GRUA	Company name: GYM	Qty: 1	Flow description	+ NEW PREDECESSOR	Flow status	Flow management action	Flow reason for varia...	Due date (mm/dd/yyyy): Date in mm/dd/y	0
WORKSPACE	Flow name (workspaces): N20 P4'			Flow description	+ NEW PREDECESSOR	Flow status	Flow management action	Flow reason for varia...	Due date (mm/dd/yyyy): Date in mm/dd/y	0

Fig. 6. Snapshot of the Activity-Flow application developed to collect activity and flow data.

3.2 Supporting Schedule and Resource Allocation Decisions on Site

One of the main production control tasks for field managers is to anticipate which activities are likely to be delayed and to implement corrective actions to prevent those delays from occurring. This section describes how the Activity-Flow Model (AFM) representation supports the generation of predictive models that allow field managers to anticipate variations in downstream activities. Predictive models require "developing a mathematical tool or model that generates an accurate prediction" (Kuhn and Johnson 2013).

To help field managers during production control, the predictive models need to provide timely and accurate predictions. Predictions about activity variations need to be made with enough time in advance to enable field manages to take corrective actions. Similarly, predictions need to be accurate to be believable and to support field managers in controlling production.

There are two types of predictions that can help field managers to proactively prevent variations during production control. The first are predictions identifying which activities are likely to be delayed. The second are predictions estimating the amount of the activity variations, which can be either negative (indicating delay) or positive (indicating early completion). Predictions identifying which activities are likely to be delayed help to alert field managers, but do not allow them to size the corrective actions. To size the corrective actions, such as choosing a time buffer, field managers need an estimate of the expected activity variations. However, generating timely and accurate predictions that estimate the variation amounts is more challenging than predicting delay likelihood because of the precision needed for the estimates. Existing methods for estimating the likelihood of activity delay and estimating the amount of activity variations require field managers to make subjective risk assessments about an activity's risk and do not leverage the flow performance record or the current flow status to make the estimates.

After a summary description of the variables collected by the AFM, we will present the development of two predictive models that estimate the likelihood of activity delay and the activity variation.

3.3 Variables Collected by the AFM

The first step towards building a predictive model is collecting the data required to build and test it. Predictive model accuracy generally improves if the data collected includes relevant variables that serve as predictors for the underlying process being modelled (Kuhn and Johnson 2013). The AFM application collects activity and flow data that has been theoretically linked to explaining activity and flow variations. Stored in databases these data become the inputs for predictive models.

The variables collected by the AFM can be classified into four groups: activity representation, activity status, flow representation, and flow status (Table 1). The AFM collects a total of 53 variables per activity and up to 46 variables per flow (the exact number of variables depends on the type of flow). Most of the variables collected are automatically calculated by the AFM, which reduces the time commitment required from field managers to implement the AFM on site. Field managers only need to enter

Table 1. Types and number of variables collected by the AFM.

	Activity	Flow
Representation	5	6
Status	48	40
Total	53	46

16 of the 99 variables collected to represent the activities and flows, the AFM automatically calculates the rest. Specifically, they need to enter six variables to create an activity and three variables to update an activity. They need to enter five variables to create a flow and two variables to update a flow.

The following sections describe the activity-level variables and the flow-level variables collected by the AFM. They also discuss whether the variables can be used for prediction, which depends on two conditions. First, the variable must be available at the time of prediction, meaning at the time the plan for the next week is elaborated. This condition tests whether the predictive model can make timely predictions by anticipating variations in the activities for the coming week without using information leading up to the activities' execution. Second, the variable's value type must be either numeric or categorical.

Activity-Level Variables

The activity-level data can be classified into two groups: representation and status. The representation variables are set when the activity is created and do not change during the use cycle of the activity. The status variables change as the activity progresses through its use cycle from creation to completion.

Activity Representation Variables

There are five variables that represent an activity: activity name, activity type, activity responsible stakeholder, activity company, and UniFormat level 4 (CSI 2018). There are two theoretical justifications for collecting these variables. The first is that the activity name, stakeholder responsible for the activity, and activity company represent a high-level <CAR> tuple, describing what action is performed on what component by what resources (Darwiche et al. 1988). The second justification comes from the Construction Method Model, where Construction Method Model Templates are associated to an application domain (Aalami 1998). Therefore, activities belonging to the same activity type or Uniformat classification are expected to behave similarly.

The activity name is unique to an activity. An activity belongs to an activity type. For example, the activity "Pour columns A1" belongs to the activity type "Pour columns." The variable activity responsible stakeholder specifies the type of stakeholder responsible for executing the activity (e.g., structural subcontractor) and the company that is responsible for executing the activity. Of these variables, only the UniFormat level 4 does not need to be specified by the field manager, since the UniFormat level 4 classification is associated with the activity type in the Activity-Flow formalization. The only variable that cannot be used for prediction is the activity name because it is unique to an activity.

Activity Status Variables

Activity status variables describe the activity's state at any point in time. The activity's status is represented by 48 variables and can be classified into four groups: activity variation metrics (27 variables, e.g., difference between the planned start and actual start), Last Planner System variables (7 variables, e.g., activity was anticipated), work-in-progress variables (6 variables, e.g., number of planned activities on a certain date), and activity risk variables (8 variables, e.g., various risk indices, see below). Field managers only need to enter data for the planned start, planned finish, actual start, actual finish, and reason for variation. The planned start and planned finish are entered during set-up, and the actual start, actual finish, and reason for variation are updated during tracking. All the other 43 activity status variables are calculated automatically by the AFM.

The first group of variables describes the status of the activity with respect to variation metrics. The second group of variables is related to the Last Planner System, characterizing the weekly plan, the activity non-completion, the reason for variation, and the tasks not anticipated. The third group of variables is related to the amount of work-in-progress going on at the construction site. A large amount of work-in-progress is linked to longer cycle times (Hopp and Spearman 2011) but is also associated with higher planning reliability (Gonzalez et al. 2010, 2011). The AFM automatically calculates the amount of work-in-progress by counting the number of activities that are being executed on a given date. The last group of variables are associated with the activity risk calculations. In this case, the activity risk is a function of the current status of the activity and its flows and the activity and flow variability record for that activity type and its flow classes on the project. The AFM automatically calculates the current activity risk as well as the baseline activity risk. Only the baseline activity risk can be used for prediction. The baseline activity risk is calculated when the look-ahead plan and the weekly plan are updated. In contrast, the current activity risk is re-calculated based on the status of the predecessor activities. Therefore, the current activity risk reflects actual data not available at the time of prediction.

Flow-Level Variables

The flow-level variables can also be classified into the two groups representation and status. The representation variables are set when the flow is created. The status variables change as the flow progresses through its use cycle: from when it is created to when it is ready to be used by the activity.

Flow Representation Variables

The flow representation variables characterize the flows by setting the flow name, flow type, stakeholder responsible for a flow, flow responsible company, predecessor name, and predecessor activity type. If the predecessor name is set the flow is on-site flow, whereas if the predecessor name is empty the flow is an off-site flow. The predecessor activity type is the only variable that is automatically added by the AFM. All the flow representation variables can be used for prediction.

Flow Status Variables

There are a total of 40 flow status variables, which are divided into 5 groups: buffer calculation (2 variables), flow variation metrics (5 variables), flow risk calculation (5 variables), flow status (22 variables), and flow utilization (6 variables). Not all the variables are applicable to all the flow types. Variables related to off-site flows (e.g., due date) do not apply to flows of type workspace or precedence, which can never be of type off-site. Similarly, variables related to on-site flows (e.g., predecessor) do not apply to flows of type external, which can never be of type on-site. See Garcia-Lopez (2017) for a detailed discussion of flow characteristics. The AFM automatically calculates most of the flow status variables except the reason for flow variation and the flow due date and ready date (for off-site flows), which are, of course, user entries. Typically, the status variables can be used for prediction except if the variable is not available at the time of prediction or if the variable is a date.

Comparison of the Variables Collected by the AFM vs Other Construction Models The AFM collects significantly more variables than the two prevalent existing construction models: The Resource-loaded Critical Path Method (RCPM) (Fondahl 1961) and the Location-based Management System (LBMS) (Kenley and Seppänen 2009).

RCPM is the most common scheduling representation used today. It is embedded into commercial project management software and is deeply integrated into jobsite planning and control workflows. RCPM's representation allows field managers to represent the precedence flows fully and to represent the labor and equipment flows partially by associating the corresponding resource as an attribute of the activity. However, it does not include information about the flow source and so cannot track the flows' status or variation metrics.

LBMS leverages the line-of-balance representation and allows field managers to plan and control the project from a location point of view. The LBMS contains a more complete representation of the activities and flows compared to RCPM because it formally represents the precedence and workspace flows. Likewise, it can represent the labor flows connecting the same activity type being carried out in different locations but cannot represent the trade flows between activities of different activity types (i.e., between flow lines). Like RCPM, the LBMS can partially represent the equipment and material flows by associating them with an activity.

Each of the variables presented in the previous section was examined to determine if it could be collected by the RCPM or LBMS models. Notice that most of these variables are not currently being collected by existing software built on the basis of these two models. We examined whether a construction model enables the collection of a variable, not whether the variable is collected by existing software.

The AFM collects 183 more variables than the LBMS and 260 more variables than the RCPM. The biggest differences are in the ability of the AFM to represent flows that originate off site, which neither the RCPM nor the LBMS can handle. Similarly, neither the RCPM nor the LMBS track the reasons for variation, nor activity and flow risks. Finally, neither the RCPM nor the LMBS represent or track information flows.

3.4 Predictive Models Enabled by the AFM

The main objective of developing the AFM was to support data-driven methods to help field managers make better and more consistent decisions in performing their look-ahead and weekly planning and control tasks and to create a basis for understanding the typical variability of types of activities. An important intuition was that more complete representation and control of the construction on-site work would yield a richer dataset that enables better predictive models.

Basic Approach to Models Predicting the Likelihood and Extent of Delays
Early successes of implementing artificial intelligence for construction control in activity-based representations provided a starting point for the AFM. For example, the Platform I knowledge expert system (Levitt and Kunz 1985) used a CPM schedule representation adding risk factors to the activities. As a project progressed and the activities were completed, Platform I identified the risk factors associated with activities with shorter durations and the risk factors associated with longer durations. The factors associated with shorter durations were classified as "Knights" while the ones associated with longer durations were classified as "Villains." Platform I then re-estimated the durations of downstream activities by increasing the planned duration of activities associated with "Villains" and reducing the planned duration of activities associated with "Knights." The underlying assumption driving Platform I was that the activities in a project are not independent but rather highly correlated based on the risk factors they share. Platform I leverages these correlations to adjust the durations of downstream activities based on the past performance of similarly correlated upstream activities.

Similarly, an assumption underlying the AFM is that activities are correlated via the flows they share. For example, the activity "Pour column" is correlated with the activity "Pour slab" because both need the labor flow "concrete crew" to be executed. If the flow "concrete crew" has failed to be ready for previous activities, it is more likely to fail in the future. Similarly, activities belonging to the same activity type are likely to be correlated because they need similar flows. Hence, the AFM can associate flows with low readiness variation with "Knights" and flows with high readiness variation with "Villains."

Garcia (2017) developed two types of predictive models. The first is a risk-based model that calculates the activity's delay risk based on the activity's variability record, the current status of the flows, and the current status of the activity. The advantage of risk models is that they do not require data to be trained. Hence, they can be implemented quickly on jobsites, which enables prospective as well as retrospective validation. A disadvantage of risk models is that they do not learn from previous examples and the predictive model is relatively simple. The second predictive model is a machine learning model that uses the Random Forest Algorithm (Breiman 2001) to predict variations of downstream activities. The advantage of machine learning algorithms is that they learn based on previous examples and can handle very complex predictive models involving many variables. A disadvantage of machine learning models is that large amounts of data are needed to train them.

Activity and Flow Risk Indices

The activity risk index is an indicator for the likelihood that an activity's start will be delayed. The flow risk index is an indicator of the likelihood of a flow not being ready. Activity and flow indices were developed based on risk models.

Risk models have been widely used in construction to identify schedule risk factors and estimate their impact on downstream activities (Akintoye and MacLeod 1997;

Chapman 1990; Dawood 1998; Liu 2010; Tah et al. 1993; Taroun 2014). A drawback of existing models is that activity risk factors and the correlations between activities need to be specified by the modeler. This has limited the practical applicability of risk models. In contrast, activity and flow risk indices rely on data that are readily collected by the AFM, and the correlations between the activities are already represented in the schedule via the flows that join them.

Risk models attempt to estimate the potential impact of an event occurring (Kähkönen and Artto 1997). Risk can be estimated by multiplying the probability of the event occurring by its expected impact.

Flow Risk Index

Field managers need to proactively manage the flows to ensure that they are ready at an activity's planned start. To do this, they need to identify the flows at risk of not being ready. The flow risk index (FRI) is a measure of the likelihood of a flow not being ready on the planned start for the activity. The flow risk index is calculated by multiplying the probability of the flow not being ready times the expected impact of the flow not being ready. The probability of the flow not being ready is given by the flow failure percentage, which is the percentage of times that the flow class (e.g., concrete crew, steel delivery) has not been ready on time. The expected impact of the flow not being ready is equal to the expected amount the flow would push the activity if it is not ready, which is calculated by comparing the expected flow delta ready with the flow buffer (Fig. 7). The expected flow delta ready is calculated by adding the total cumulative delay up to the activity releasing the flow and the mean delta ready for the flow. Figure 7 also illustrates the calculation of the total delay accumulated up to the activity releasing the flow, the mean flow delta ready, and the flow buffer. On the other hand, if the flow is ready (i.e., has been delivered on site or has been released by the upstream activity) the flow risk impact is calculated as the flow delta ready minus the flow buffer. Notice that the dimension of the flow risk index is in days. It can be interpreted as the estimated number of days that a flow will push the successor activity.

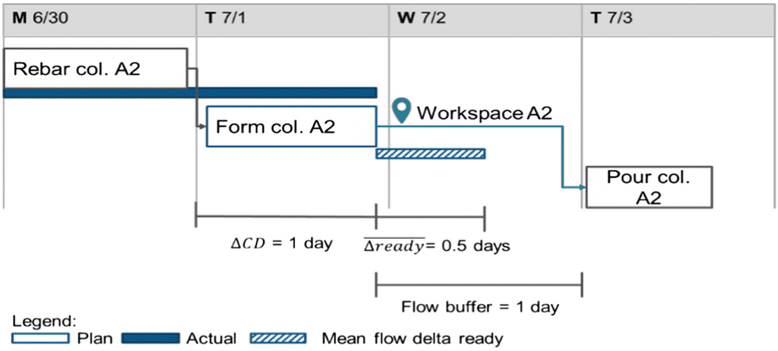


Fig. 7. Example showing calculation for total delay accumulated up to an activity

Activity Risk Index

Field managers also need to identify the activities at risk of delay. The activity risk index (ARI) is a measure for the likelihood that an activity will not start on time. It is calculated based on the current risk of the flows feeding it, the activity variability record, and the status of the activity. The activity risk index is calculated by multiplying the probability of the activity starting late times the expected impact of the delay. The probability that the activity will start late is given by the percentage of times that the activity type has started late. The impact of the activity not starting on time is given by the mean delta start for that activity type times the maximum flow risk index feeding the activity. Figure 8 shows an example for calculating the maximum flow risk index and the activity type's mean delta start. Finally, the activity status index is used to scale the risk according to the current status of the activity. The activity status index is set to 1 if the activity has not started, 0.5 if it has started, and 0 if it has finished. Notice that the dimension of the activity risk index is days squared, resulting from multiplying the mean delta start (days) by the maximum flow risk index (days). This means that the activity risk index does not have a readily comparable physical interpretation.

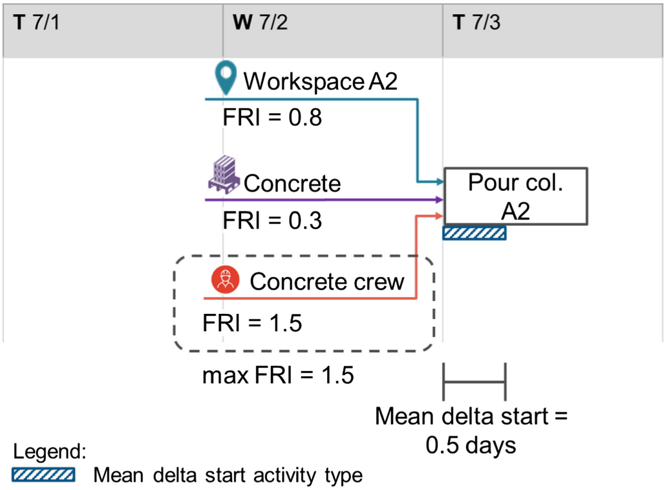
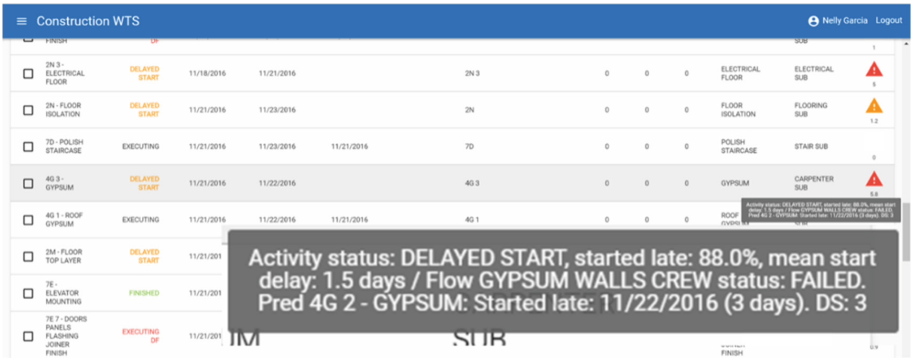


Fig. 8. Example showing the calculation of the maximum flow risk index and the mean delta start.

Results Implementing the Activity and Flow Risk Indices on Test Projects

The activity and flow indices were tested by implementing them on three test projects and validating them prospectively and retrospectively. The risk indices were developed during the initial eight-week data collection period on project A. Therefore, the validation period for project A occurred during the last ten weeks of data collection. The validation for the projects B and C occurred during the last three weeks of the data collection because one week was needed to collect the data to generate the risk predictions.

The prospective validation was carried out by adding activity and flow alerts to the Activity-Flow application that helped the field managers identify the activities and flows at risk of being delayed. Once the activity and flow risk indices had been calculated for all the activities and flows in the plan, the AFM assigned a risk classification according to the risk indices' value. The risk indices whose value was in the upper 75% of the project were classified as "High," those between the 50%–75% were classified as "Medium" and those in the lower 50% were classified as "Low." The risk classification was used to alert the field managers visually in the Activity-Flow application. Activities and flows with a high-risk classification had a red alert, those with a medium-risk had an orange alert, and those with a low-risk classification had no alert. The activity flow alert was shown in the main table view for the activities (Fig. 9), while the flow risk alert was shown in the activity detail view. A popup describing why the activity or flow was at risk appeared when the field managers hovered over the alert. This popup was added based on feedback from the field managers who wanted to understand how the application made its assessments.



☐	2N 3 - ELECTRICAL FLOOR	DELAYED START	11/18/2016	11/21/2016		2N 3	0	0	0	ELECTRICAL FLOOR	ELECTRICAL SUB	▲ 5
☐	2N - FLOOR ISOLATION	DELAYED START	11/21/2016	11/23/2016		2N	0	0	0	FLOOR ISOLATION	FLOORING SUB	▲ 1.2
☐	7D - POLISH STAIRCASE	EXECUTING	11/21/2016	11/23/2016	11/21/2016	7D	0	0	0	POLISH STAIRCASE	STAIR SUB	0
☐	4G 3 - GYPSUM	DELAYED START	11/21/2016	11/22/2016		4G 3	0	0	0	GYPSUM	CARPENTER SUB	▲ 5.8
☐	4G 1 - ROOF GYPSUM	EXECUTING	11/21/2016	11/22/2016	11/21/2016	4G 1	0	0	0	ROOF GYPSUM		
☐	2M - FLOOR TOP LAYER	DELAYED START	11/21/201									
☐	7E - ELEVATOR MOUNTING	FINISHED	11/21/201									
☐	7E 7 - DOORS PANELS FLASHING JOINER FINISH	EXECUTING DF	11/21/201							FINISH		0.9

Activity status: DELAYED START, started late: 88.0%, mean start delay: 1.5 days / Flow GYPSUM WALLS CREW status: FAILED. Pred 4G 2 - GYPSUM: Started late: 11/22/2016 (3 days). DS: 3

Fig. 9. Snapshot of the Activity-Flow application showing the activity alerts.

The field managers on the three test projects provided feedback about the alerts. They mentioned that the alerts helped them communicate about issues with the subcontractors, such as the impact that they were having on downstream activities. For example, on project C, the superintendent showed the carpentry crew that they were delaying the flooring crew because they had not released some of the workspaces on time.

Another example occurred on project A. The Activity-Flow application alerted the field managers about a possible workspace conflict resulting from the interference between the self-climbing scaffold and the level 19 slab pour. The self-climbing scaffold released the space that was needed for pouring the slab. There had not been a workspace interference between these two activities before because the self-climbing scaffold had always been a few levels ahead of the slab. Once the self-climbing scaffold reached its final level, the supplier was supposed to remove it from the jobsite. The field manager responsible for the self-climbing scope arranged the removal on July 20. However, the field manager responsible for the tower shell construction had planned the start of the level 19 slab pour on July 18. Since the look-ahead planning for project A was done using Excel sheets and each field manager elaborated the look-ahead for just their own scope, this conflict had not been spotted. Once the alert was spotted and validated, the field manager responsible for the self-climbing scope renegotiated the date for the removal of the self-climbing scaffold to start on July 16, avoiding the workspace conflict that would have stopped the construction of the main tower from progressing.

A retrospective validation of the activity risk indices was carried out at the end of the data collection period. The objective was to determine whether the activity risk index was positively correlated with the activity start variation. To achieve this, the AFM application saved the activity risk indices calculated when the weekly plan was set. This meant that if the weekly plan was elaborated on a Friday, the AFM application would calculate and save the risks for the activities in the weekly plan for the next week. Since field planners are supposed to commit to activities that they can carry out (per the Last Planner System), they should only commit to activities with a low risk or a risk they can mitigate during the week. If the AFM application did not save the activity risk indices but rather took the activity risk calculated at the end of the week, the predictive power of the activity risk could not be asserted because the activity risk would reflect the delays that happened during the week. At the end of the week, the AFM application calculated the activity delta start for the activities in the weekly plan.

During the week, the field managers made changes to the sequencing, priority, and resource allocations to prevent activity delays. The second author kept a record of these changes in an intervention log to identify the activities affected by the changes. However, it was difficult to discern what actions had been triggered by the activity and flow risk alerts and which were based on the field managers' experience. Additionally, it was impossible to establish the actual activity start delay if no adjustment had been made. This is a methodological limitation inherent to carrying out research on real construction projects instead of performing experiments or simulations in controlled environments. Testing the strength of the activity risk predictions against the activity start delay is biased against the activity risk predictions because field managers seek to minimize activity delays. Hence, activities with a high activity risk prediction might have been adjusted, leading to a lower activity start delay than would have been expected if the activity had not been adjusted. Therefore, a positive correlation between the predictions and the activity start delay would provide very strong evidence of the strength of the activity risk predictions.

To test the activity risk predictions, a simple linear regression between the activity risk prediction and the activity delta start was carried out for the three test projects. The linear regression fit was plotted as an overlay of the scatterplot between the activity delta start (y axis) and the activity risk prediction (x axis) (Fig. 10). Each point in the scatterplot represents an activity. Note that there are several points that can overlay each other.

The linear regressions for the three projects are statistically significant to a 0.005 level. Furthermore, the correlation between the variables is positive for the three projects, indicating that a higher activity risk prediction is correlated with a higher activity delta start. The correlation for the project C was the highest (0.55) followed by project B (0.38) and project A (0.36). However, the linear model is unable to explain

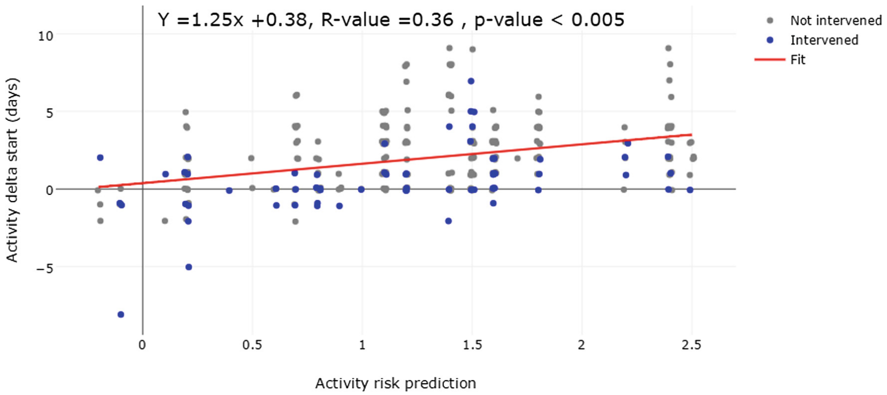


Fig. 10. Linear regression results for predicting the activity delta start using the activity risk prediction for project A.

most of the variation in the underlying data (low R-value). As discussed earlier, one of the reasons why the fit is weak is that field managers adjusted the activities during the week. Since most of the intervened points lie on the lower side of the linear fit, it can be expected that without adjustment those points would have experienced a higher variation, possibly leading to a better fit. Additionally, activities adjusted by the field managers tended to have a higher risk than those that were not adjusted. This provides evidence of the usefulness of the activity risk indices for helping field managers anticipate delays and also of the predictive power of the activity risk indexes.

3.5 Machine Learning Model

In addition to identifying the activities likely to be at risk of delay, field managers also need models that quantify the amount of the expected activity variations. The development of a machine learning model that leverages the dataset collected by the AFM to predict variations in the delta start, delta duration, and delta finish of downstream activities is described below. Machine learning is the "study and computer modelling of learning processes" (Michalski et al. 1983). Machine learning algorithms learn from data to calibrate their parameters and interrelationships to generate predictions (Kuhn and Johnson 2013). There are two main types of problems that can be solved using machine learning algorithms: supervised and unsupervised problems. Supervised learning problems are those where the data are labeled, i.e., the algorithm knows the value of the outcome variable during training. Unsupervised learning problems are those where the data are not labeled, i.e., the algorithm does not know the value of the outcome variable. The AFM-based prediction model is an example of a supervised learning problem because the dataset contains the outcome variables to be predicted: the activities' delta start, delta duration, and delta finish.

Supervised learning problems can be divided into two additional categories: regression and classification problems. The distinction between classification and regression problems depends on the nature of the outcome variable. Regression problems attempt to predict the value of the outcome variable. Classification problems attempt to classify the outcome variable into categories or groups. This research deals with a regression problem since it seeks to predict an activity's variation metrics.

Description of Dataset

Machine learning algorithms require a large amount of data to train their parameters. The only test project that had enough data available for training and testing the algorithm was project A with 18 weeks of data yielding a total of 1,153 activities, 4,192 flows, and 243,029 data-points. The Activity-Flow Model stored the activity and flow data in two databases. The two databases are linked via the activity name and the baseline date. The reason for storing the data in these two databases was that an activity does not necessarily have to have all the seven flows linked to it. Moreover, an activity can have more than one flow of the same type. For example, the activity "Build CMU wall" can have three material flows: CMU blocks, rebar, and grout. Nevertheless, most machine learning algorithms require that the data be arranged in a unique tabular format so that it can be used as input. Since the outcome variables to be predicted are related to the activity, the logical arrangement of the dataset was to have one activity per row. Each row contains a total of 234 variables, 23 related to the activities and 211 related to the seven flows.

With the dataset organized in this way, the machine learning algorithm can learn what activity types need what types and classes of flows and use the performance of the activity and flow types for prediction. A problem with this representation is that there can only be a single flow for every flow type. This problem was overcome by iterating over the flow database and selecting the flow with the highest flow delta push for each flow type and each activity, which we call "critical flow". Another problem with this representation is that it contains a very large number of variables and many of the columns are sparse. Therefore, one of the criteria for selecting a machine learning algorithm was its ability to deal with many variables and a relatively sparse representation. For these reasons, the Random Forest algorithm was selected (Breiman 2001). To test whether the AFM supports better predictions of activity and project duration the machine learning algorithm was applied to the dataset of project A but selecting only the predictive variables that the RCPM and LBMS representations support.

Random Forest Algorithm

The Random Forest algorithm works by generating many random regression trees using bootstrapped samples of the input dataset and using the output of the random regression trees to make the prediction (Breiman 2001). One of the main advantages of the Random Forest algorithm is that it has cross-validation built in. Each regression tree in the random forest is trained using a bootstrapped subset sample of the original dataset. Hence, each tree can be tested by using the Out Of Bag (OOB) data, i.e., data not in the training subset. The testing error is calculated by aggregating the OOB errors for the trees. This avoids the need to divide the dataset into training and testing sets, allowing more data to be leveraged for learning.

Nine models were run to explore all the combinations of the different outcome variables to be predicted (delta start, delta duration, or delta finish) and the variables represented by the three construction models (RCPM, LBMS, and AFM). All of the

Random Forest runs used the same number of random trees (1,500), the node size (5), and the same action for dealing with missing data. Similarly, all the runs were initialized using the same random number seed (754) to ensure reproducibility in the results and allow comparison between the construction models.

Regression Results

The AFM outperforms the RCPM and the LBMS representations for predicting all the activity variation metrics. Similarly, the LBMS outperforms the RCPM representation for predicting all the activity variation metrics. The AFM explains 11% more variation than the LBMS and 20% more variation than the RCPM representation for predicting the activity delta start. Similarly, the Activity-Flow Model explains 9% more of the variation than the LBMS and 41% more variation than the RCPM representation for predicting the delta duration. Finally, the Activity-Flow Model explains 10% more variation than the LBMS and 19% more variation than the RCPM representation for predicting the delta finish (Fig. 11). All the representations perform better at predicting the delta start and the delta finish than at predicting the delta duration. This is understandable since the models represent, to different degrees, the readiness variation of the construction flows. For instance, the RCPM can represent the precedence flow readiness variation, while the LBMS can represent the precedence and workspace flow readiness variation. However, none of the models measure the variation in the quality and quantity of the flows, which tends to affect the duration of the activity. For example, if the activity “Pour columns” has a planned labor flow consisting of a three-person concrete crew and one person becomes sick, it is likely that the activity will start on time but the duration will be longer than planned. This is a limitation of the AFM since it tracks the readiness of the flows but not the quality of the flows. While more research is needed to create better predictions of the amount of expected delays so that field managers can prioritize their interventions, these results show that AFM outperforms the other representations and that tracking flows yields a more complete representation of the on-site work and can be leveraged for developing better predictive models.

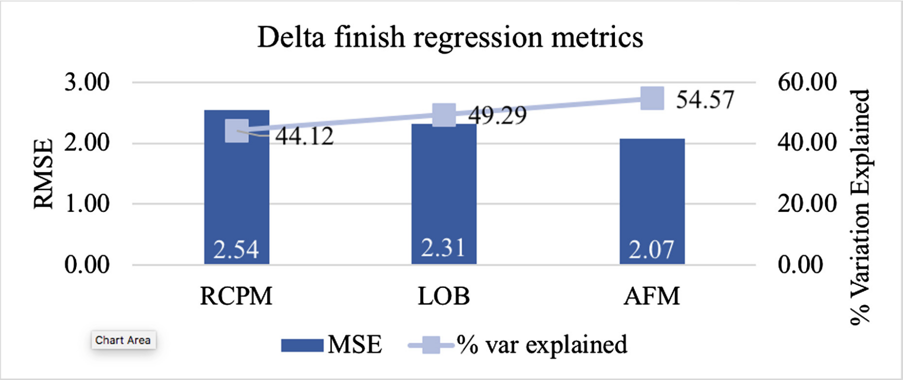


Fig. 11. Regression results for predicting the activity delta finish for the three construction models.

3.6 Summary of AFM

The AFM provides a more complete representation of the on-site work compared to existing construction models, namely, the RCPM and the LBMS representations. This more complete representation allows the AFM to support predictive models that help field managers anticipate variability in downstream activities. Two predictive models were developed that leverage the AFM representation: a risk-based model and a machine learning model. The risk-based model leverages the record of project activity and flow variability and the current status of the activities and flows to estimate the risk of an activity starting late. The activity risk index was validated on three projects by comparing the risk predictions against the actual activity start variation. The activity risk index prediction was a statistically significant predictor of activity start variation for the three projects. Additionally, it enabled alerts to be generated to warn field managers about the downstream activities and flows that are at risk of being delayed. The machine learning model leveraged the Random Forest algorithm and used the data from project A to make predictions about the activity start, activity duration, and activity finish. The models that used the AFM representation outperformed the models that used LBMS and RCPM representations for predicting the delta start, delta duration, and delta finish. Furthermore, five of the ten most important predictive variables identified by the Random Forest algorithm were related to representing and tracking the flows. Hence, representing and tracking the flows is important for generating better predictions that can help field managers anticipate variations and manage production in the field. Future work could address the quality of flows and supply chain flows so that understanding the flows directly related to activities on site is complemented by an understanding of the performance of the supply chain activities and their flows.

4 Conclusion

The two examples presented in this paper showed the application of intelligent computing to the construction scheduling domain. The TCM research was based on several prior Ph.D. theses that had deployed intelligent computing methods, such as semantic modeling and genetic algorithms (Aalami 1998; Dong 2012). For the AFM research, the failed attempt to gain insights into patterns of activities that are more prone to delays than others through the application of machine learning methods to an activity-based dataset made us realize the limitations of predicting the duration of the construction phase of a project purely on the basis of activities. The identified problems then led us on a search for applicable concepts, which we found outside construction for TCM – how to handle various types of constraints computationally – and in the construction domain for AFM – the seven types of flows conceptualized by Koskela (1999). It also led us to observe practice through a new lens – space utilization during construction for TCM and the impact of misaligned flows for AFM. Combining the new concepts with the deeper understanding of the problem to be addressed provided the foundation to conceptualize a new underlying representation for construction schedule information and algorithms that leverage the new representation, TCM and AFM respectively, which, in turn, became the basis for new software tools. These software tools then allowed practitioners to deal with phenomena that are important for their projects but could not be handled before, such as the different types of constraints, the large effort to create one schedule, the difficulty in knowing whether the schedule produced is good, and the network of flows. This allowed the researchers to test the new conceptualization that embodies more fundamental knowledge in practice and collect data consistently and rapidly through simulation and rapid feedback from experienced professionals for TCM and by deploying the AFM on site to support daily and weekly planning. The consistent data then allowed the application of machine learning methods to detect patterns and make predictions that could not be done in datasets based on the previous representations.

There is, of course, still much work to do until a "truly" intelligent scheduling method is in place. The automated optimization of schedules is one such area (today's conceptualization of the TCM requires manual changes to resource availability and other constraints). The consideration of the quality of flows and of capacity and inventory buffers is another such area, in this case, related to the AFM. Thanks to intelligent computing methods, it is now, however, imaginable to embark on research in these areas.

Acknowledgments. The authors thank the members of the Center for Integrated Facility Engineering for the support of the research reported in this paper. They also thank Cynthia Brosque for her input to the manuscript.

References

Aalami, F.: Using construction method models to generate four-dimensional production models. Ph.D. Dissertation, Stanford University (1998)

Akintoye, A.S., MacLeod, M.J.: Risk analysis and management in construction. Int. J. Proj. Manage. **15**(1), 31–38 (1997)

Ballard, G., Howell, G.: Toward Construction JIT. Lean Construction, Luis Alarcón, Editor, Balkema, pp. 304–312 (1995)

Ballard, G., Howell, G.: Shielding production: essential step in production control. J. Constr. Eng. Manage. **124**(1), 11–17 (1998)

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

Chapman, C.B.: A risk engineering approach to project risk management. Int. J. Proj. Manage. **8** (1), 5–16 (1990)

CSI (Construction Specifications Institute) (2018). UniFormat. https://www.csiresources.org/practice/standards/uniformat. Accessed 15 Feb 2018

Darwiche, A., Levitt, R., Hayes-Roth, B.: OARPLAN: generating project plans by reasoning about objects, actions, resources. Artif. Intel. Eng. Des. Anal. Manuf. **2**(3), 169–181 (1988)

Dawood, N.: Estimating project and activity duration: a risk management approach using network analysis. Constr. Manage. Econ. **16**(1), 41–48 (1998)

Dong, N.: Automated look-ahead schedule generation and optimization for the finishing phase of complex construction projects. Ph.D. thesis, Stanford University (2012)

Fondahl, J.W.: A non-computer approach to the critical path method for the construction industry. Technical report No. 9, Construction Institute, Stanford University (1961)

Garcia-Lopez, N.P.: An activity and flow-based construction model for managing on-site work. Ph.D. thesis, Stanford University (2017)

Garcia-Lopez, N.P., Fischer, M.: A construction workflow model for analyzing the impact of in-project variability. ASCE Construction Research Congress, 31 May–2 June, San Juan, Puerto Rico (2016)

Gonzalez, V., Alarcón, L.F., Maturana, S., Mundaca, F., Bustamante, J.: Improving planning reliability and project performance using the Reliable Commitment Model. J. Constr. Eng. Manage. **136**(10), 1129–1139 (2010)

Gonzalez, V., Alarcón, L.F., Maturana, S., Bustamante, J.A.: Site management of work-in-process buffers to enhance project performance using the Reliable Commitment Model: case study. J. Constr. Eng. Manage. **137**(9), 707–715 (2011)

Hopp, W.J., Spearman, M.L.: Factory Physics. Waveland Press (2011)

Kenley, R., Seppänen, O.: Location-based Management for Construction: Planning, Scheduling and Control. Routledge (2006)

Kenley, R., Seppänen, O.: Location-Based Management for Construction: Planning, Scheduling and Control. Spon Press, Abingdon (2009)

Koskela, L.J.: Management of production in construction: a theoretical view. In: 7th Annual Conference of the International Group for Lean Construction, Berkeley, CA, pp. 241–252 (1999)

Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013)

Levitt, R.E., Kunz, J.C.: Using knowledge of construction and project management for automated schedule updating. Proj. Manage. J. **16**(5), 57–76 (1985)

Liu, J.: Bayesian network inference on risks of construction schedule-cost. In: 2010 International Conference of Information Science and Management Engineering, pp. 15–18. IEEE (2010)

Michalski, R.S., Carbonell, J.G., Mitchell, T.M.: Machine Learning: An Artificial Intelligence Approach. Springer, Heidelberg (1983)

Morkos, R.: Operational Efficiency Frontier: Visualizing, Manipulating, and Navigating the Construction Scheduling State Space with Precedence, Discrete, and Disjunctive Constraints. Ph.D. thesis, Stanford University (2014)

Riley, D., Sanvido, V.: Patterns of construction-space use in multistory buildings. J. Constr. Eng. Manage. **121**(4), 464–473 (1995)

Tah, J.H.M., Thorpe, A., McCaffer, R.: Contractor project risks contingency allocation using linguistic approximation. Comput. Syst. Eng. **4**(2–3), 281–293 (1993)

Taroun, A.: Towards a better modelling and assessment of construction risk: insights from a literature review. Int. J. Proj. Manage. **32**(1), 101–115 (2014)

Thomas, H.R., Riley, D.R., Sinha, S.K.: Fundamental principles for avoiding congested work areas—a case study. Pract. Periodical Struct. Des. Constr. **11**(4), 197–205 (2006)

Williams, J.M.: Project Production Management. Presentation in Virtual Design and Construction course, CIFE, Stanford, 24 January (2018)

Waugh, L.M.: A construction planner. Ph.D. thesis, Stanford University, Stanford, USA (1990)

Fiatech: History and Perspectives on Future Research

William J. O'Brien(✉) and Vineeth Dharmapalan

The University of Texas at Austin, Austin, TX 78712, USA
wjob@mail.utexas.edu, vineeth.dp@utexas.edu

Abstract. Fiatech, an industry-funded research consortium to promote the development and adoption of innovative technologies for the capital project industry, has played an important role in the industry. This paper presents a brief history for Fiatech since its inception in 2000 until its merger with the Construction Industry Institute in 2018. The paper presents the perspective of the authors who have been active participants and leaders within the Fiatech community. Early contributions of Fiatech include the Capital Projects Technology Roadmap and research streams in interoperability, technology evaluation, and regulatory streamlining. More recent developments include a reinvigoration of Fiatech and its roadmap through Productivity Advancement Targets, which focus efforts on adding business value and curating existing knowledge as well as identifying development needs. A supporting business case and selected enablers and project recommendations are reviewed. Implications of this effort for the research community are briefly discussed.

Keywords: Interoperability · Technology evaluation · Automated code checking
Digital seals · Construction supply chain · Construction productivity

1 Introduction

On January 1, 2018, Fiatech officially merged with the Construction Industry Institute to combine efforts and pursue significant (some would say radical) efforts to dramatically improve the performance of the capital projects industry. From 2000-2017, Fiatech made a significant impact on the research, development, and deployment of information technology in support of projects. This paper presents a brief history of Fiatech, particularly from a research perspective to document some of the key contributions and also to report to the community current perspectives of Fiatech in terms of next steps for research and development.

This paper presents Fiatech from the perspective of the authors. The first author participated since Fiatech's inception as a contributor to its technology roadmap, champion and leader for its roadmap committee, Board member, and, finally, Transition Manager in the final year of independent operation. The second author was the last graduate research assistant funded by Fiatech and participated in direct support of its most recent efforts. The paper begins with a short history of Fiatech including a review of its most significant research streams. Recent developments that focus efforts on

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 296–310, 2018.
https://doi.org/10.1007/978-3-319-91635-4_16

productivity improvements are reviewed in more detail, including a selection of enablers and projects identified by industry subject matter experts. The paper concludes with perspectives on the future as well as brief thoughts on implications for research and development efforts by the academic community.

2 History of Fiatech

Fiatech represents a focused effort by industry – in particular, the large capital projects industry – to grapple with the opportunities and challenges of computer integrated construction. This section reviews the precursor efforts that led to the formation of Fiatech, the initial efforts of Fiatech at its founding, research streams as a mature organization, and reinvigoration in recent years with targeted efforts to enhance productivity and decrease cost on capital projects.

2.1 Beginnings – CII and FIAPP

Fiatech is a creation of the Construction Industry Institute (CII). CII was founded as an organized research unit of The University of Texas at Austin in 1983 with the mission of making a measurable improvement to the performance of capital projects. Funded by its members (large owner and engineer/contractor organizations), CII developed a robust history of conducting research and supporting academics from many universities. CII has made many advancements in knowledge and practice. Notable accomplishments include significant enhancements to the practice of safety [1] and front-end planning [2]. As a member-driven organization, CII conducted research on a variety of topics of interest to its constituents. While not necessarily known for technology-related research, CII has long been interested in significant breakthroughs and conducted research in technology topics since the 1980s, the first in CAD for design [3] and cost-effectiveness of computerization [4].

Growing awareness of the potential of information technologies led CII to adopt as a concept for advancement a Fully Integrated and Automated Project Process (FIAPP) [5], sponsored by CII's Breakthrough Strategy Committee. It was understood at the time that there were many barriers holding back adoption of FIAPP concepts and related technologies. As noted on the Fiatech website, "In 1999, the Breakthrough Strategy Committee identified that efforts to realize the breakthrough promise of FIAPP in the construction industry had produced only modest progress due to the character of the industry: highly fragmented, project-oriented, multiple stakeholders, and low R&D investment. They also noted at the time that the approach of FIAPP was sporadic, independent and lacking critical mass. Worse yet, progress was stymied by a lack of common standards and protocols and by the inability to integrate software and systems improvements effectively" [6].

With a perceived need for more focused efforts to overcome barriers to adoption, Fiatech was founded in 2000 as a separate organization from CII. Sponsored by CII and in partnership with the National Institute of Standards and Technology (NIST), Fiatech was created to bring focus and accelerate the development and deployment of FIAPP in the capital projects industry. Fiatech was a member-funded organization comprised of a subset of CII member companies, technology solution providers, and members of the Owner-Operator Forum (OOF). The OOF was founded in 1999 by The Dow Chemical Company, DuPont, Air Products & Chemicals, BASF and Merck & Co. with the goal of communicating software requirements related to plant life cycle activities. With overlap in membership, the OOF joined Fiatech, and a single organization was created with the perceived breadth and depth of membership to make an impact on technology development and adoption.

2.2 Initiation - the Capital Projects Technology Roadmap

The Fiatech name results from the concatenation of "fully integrated and automated technologies." With the charge at its founding to accelerate the development and deployment of FIAPP for capital projects, initial efforts went towards the development of a plan to guide investment. The result was the creation of the Fiatech Capital Projects Technology Roadmap (CPTR) [7]. The CPTR was created over the first few years of Fiatech's existence, with authorship distributed across multiple teams of passionate experts from industry, government, and academia volunteering their time to the effort. The initial definition of the CPTR was developed in a series of workshops that identified nine central elements and their interconnections; there are five elements specific to traditional characterization of project phases and four cross-cutting elements that support the whole. The elements are:

1. Scenario-Based Project Planning
2. Automated Design
3. Integrated and Automated Procurement and Supply Network
4. Intelligent and Automated Construction Job Site
5. Intelligent, Self-Maintaining, and Repairing Operational Facilities
6. Real-time Project and Facility Management, Coordination, and Control
7. New Materials, Methods, Products, and Equipment
8. Technology- and Knowledge-Enabled Workforce
9. Lifecycle Data Management and Information Integration.

The CPTR is represented in Fig. 1, an image that for many years was an iconic representation of Fiatech and one that resonated with practitioners to help them make sense of what FIAPP signified for the industry. Each CPTR element was defined by a team of volunteers with expertise in the area. Overall the CPTR was maintained in an evergreen state by a committee with an overall leadership team and leaders attached to each element. Volunteer leaders were facilitated by a small team of Fiatech staff. Volunteers have rotated over time, but many have been consistent contributors for much of Fiatech's existence.

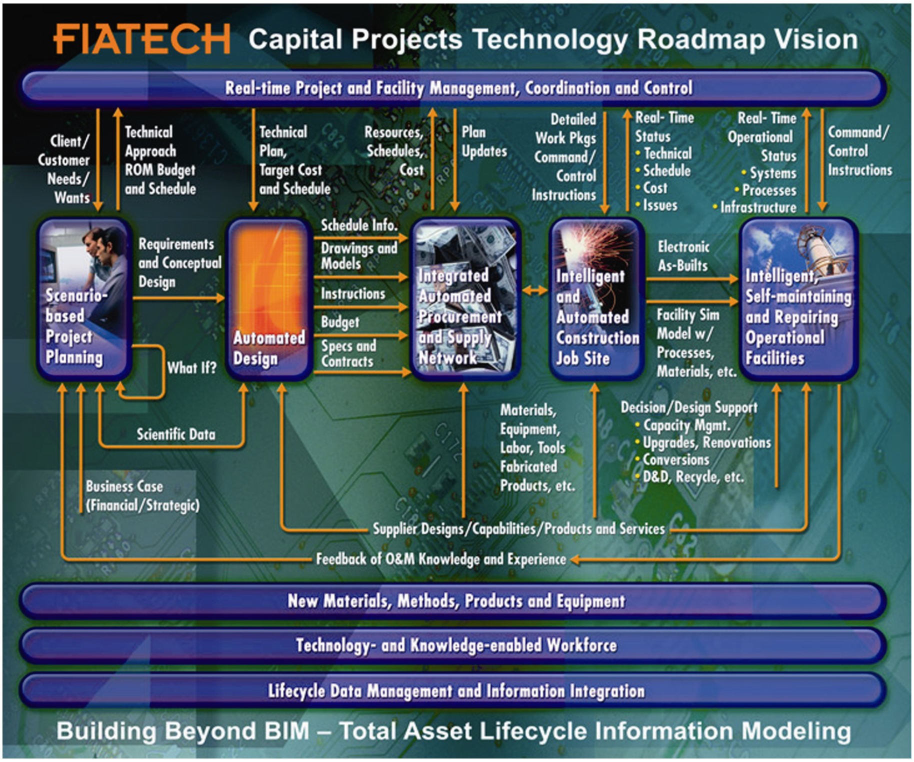

FIATECH Capital Projects Technology Roadmap Vision
Real-time Project and Facility Management, Coordination and Control
Client/ Customer Needs/ Wants
Technical Approach ROM Budget and Schedule
Technical Plan, Target Cost and Schedule
Resources, Schedules, Cost
Plan Updates
Detailed Work Pkgs Command/ Control Instructions
Real- Time Status • Technical • Schedule • Cost • Issues
Real- Time Operational Status • Systems • Processes • Infrastructure
Command/ Control Instructions
Scenario-based Project Planning
Requirements and Conceptual Design
What If?
Automated Design
Schedule Info.
Drawings and Models
Instructions
Budget
Specs and Contracts
Integrated Automated Procurement and Supply Network
Intelligent and Automated Construction Job Site
Electronic As-Builts
Facility Sim Model w/ Processes, Materials, etc.
Intelligent, Self-maintaining and Repairing Operational Facilities
Scientific Data
Business Case (Financial/Strategic)
Materials, Equipment, Labor, Tools Fabricated Products, etc.
Decision/Design Support • Capacity Mgmt. • Upgrades, Renovations • Conversions • D&D, Recycle, etc.
Supplier Designs/Capabilities/Products and Services
Feedback of O&M Knowledge and Experience
New Materials, Methods, Products and Equipment
Technology- and Knowledge-enabled Workforce
Lifecycle Data Management and Information Integration
Building Beyond BIM – Total Asset Lifecycle Information Modeling

Fig. 1. The Fiatech Capital Projects Technology Roadmap vision for an integrated and automated project processes with supporting technologies. (from [8])

The Fiatech elements represent a broad and encompassing vision for the capital projects industry; the related roadmap documents constituted several hundred pages of project plans. The first author recalls conversations that priced the roadmap at 100 million USD – an imprecise sum but one that captures the scale and scope of the aspiration to change the industry. Some sense of the scale and detail of the CPTR documentation can be seen in Fig. 2, which represents the initial plan for the procurement and supply network element [7].

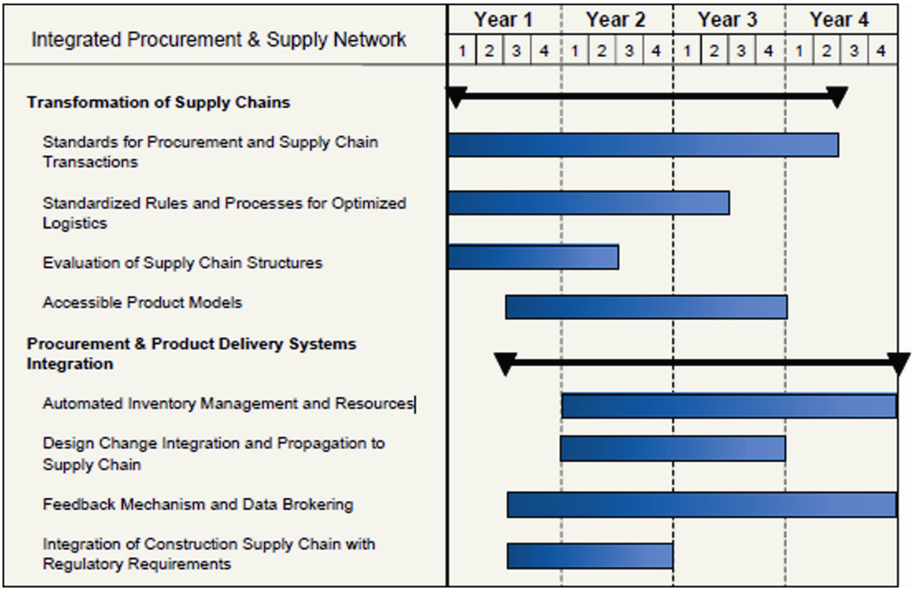


Fig. 2. Initial CPTR plan for the integrated procurement and supply network element, listing project titles and schedule. (from [7])

2.3 Maturity and Research Streams

Building from the early work on roadmap development, Fiatech was chartered to conduct research and development projects from the roadmap. In one sense Fiatech was very successful in that the roadmap elements and broader vision helped to catalyze various efforts in academia, industry, and government. Many academics participated in the initial development of the CPTR and many more referenced it when developing proposals. To some extent, government agencies aligned with the roadmap – in particular, NIST, which was a significant participant in Fiatech. Perhaps most important, the roadmap vision was one that translated well to industry and was used to inspire many corporate efforts and thus in its way influenced many millions of dollars of development.

Despite its influence, Fiatech was never as successful as hoped in terms of raising funds and directing and coordinating research and development efforts. Thus the plans and timing of the CPTR never were fully realized. A reading of the projects on the supply network listed in Fig. 2 shows that many of the topics are still relevant today (see discussion below). With limited resources, Fiatech over its life did conduct research on several topics in some broad streams. Notable streams include information integration, evaluation activities, and regulatory streamlining.

Support for information integration, in particular, development of interoperability data standards and supporting infrastructure, was a long-term focus of Fiatech. In this, it cooperated with several other standards groups such as MIMOSA, POSC Caesar, and DEXPI as well as conducting independent development projects. Much of this work was in support of advancing the ISO 15926 data standard for process plants [9]. At least half of Fiatech's investment in research and development over the years was in support of information integration standards, with nearly 30 project deliverables (examples include [10–13]). While much of this work was conducted in concert with other organizations, it is fair to say that Fiatech has had a significant footprint advancing the state of practice in interoperability for the process industry.

A second research stream was in the area of evaluation of technology implementations. Such work broadly involved documenting benefits, but also lessons learned from implementations to speed adoption and refine the technologies. An early success was the "smart chips" series of projects on using field location technologies for productivity, such as for tracking pipe [14, 15] and for tools [16]. Later work in evaluation in this area has focused on how people utilize smart tools through surveys [17] and direct observation [18]. Of note, much of this work was conducted with the academic community and resulted in academic publications [19–21]. Some related work has focused on documenting information needs to support implementation, such as a survey on operations and maintenance needs [22].The area of evaluative research has had high impact. The early work on smart chips played a role in helping the industry evaluate lessons learned and speed adoption of RFID and related technologies for materials tracking. Technology evaluation was perhaps an area that Fiatech could have invested more; only about 15 studies were published in this area or about half of those in the information integration area. This is perhaps due to some overlap with CII, which in the same years was conducting research on benchmarking and evaluating the use of information technologies [23–25] and even making recommendations about technology adoption [26].

More focused but long-term research and development have been in the area of regulatory streamlining. Unlike other areas of Fiatech interest that have been led by large capital projects organizations in the process sector, these projects were led by the medical and retail members of Fiatech who build many smaller projects and must meet building standards and permitting processes in multiple jurisdictions. There have been two parallel activities in this area. "AutoCodes" has the express goal of reducing plan review from weeks and months to hours and days. Development here has taken place in three stages; stage one demonstrated the feasibility of automated code checking [27], phase two developed model protocols, expanded rule sets and building types, and developed outreach materials to speed electronic plan review [28]. Phase three is on-going with an effort to broaden deployment. The other activities are in digital seals and signatures, where Fiatech has undertaken survey research to understand the opportunities and challenges in moving from wet seals and signatures to electronic documents [29]. On-going work seeks to further the recommendations for best practices to obtain broad deployment of electronic seals and move away from paper (or keeping an electronic process electronic).

The three themes reviewed above represent the principal areas of research and development sponsored directly by Fiatech. Numerous other areas were investigated; some notable ones include workforce skills [30], supporting NIST projects on guidance for capital facilities and building information handover [31, 32], automated design [33, 34], and some data specifications to support purchasing of equipment [35]. All of Fiatech reports are published on the Fiatech website; with the transition to CII, reports are being moved to the CII Knowledge Base website which should give them visibility to a broader audience.

2.4 Recent Developments

For an extended period, the Capital Projects Technology Roadmap was an organizing meme for how Fiatech has looked at projects. The CPTR was initially used to identify projects early in Fiatech's existence. As Fiatech matured and found it did not have the resources to fund the CPTR fully, it still had a road map committee and champions who identified and vetted projects. Projects needed to identify closely with an individual element or clearly sit within a few elements. This process served Fiatech well, although for the past several years the roadmap has been seen as confining as many projects of interest to members were seen to span multiple elements. In many ways, this challenge is due to the success of the CPTR vision – when first presented, the vision as depicted in Fig. 1 was novel for the industry. As the industry matured and adopted much of the vision, the CPTR was no longer as compelling as it was during the early days of Fiatech.

In recent years, Fiatech members have sought a more focused effort to accelerate the fulfillment of the CPTR vision. With that, there has been a focus on achieving meaningful results for the business, in particular by increasing productivity and decreasing costs for capital projects. Part of this desire was driven by owner organizations who are finding the stagnating productivity of the construction industry is making capital projects too expensive to undertake. The desire for more focused efforts centered on business results resulted in the development of a set of target goals broadly aligned with contemporary project execution approaches. There were designated Productivity Advancement Targets (PATs), each with a goal and a team of experts behind them. The PAT teams in the past two years have formed, curated existing knowledge in the area, identified enablers to help firms achieve success with existing technologies, and identified gaps that need to be addressed with projects. These developments have taken most of the effort of Fiatech members in the last few years and represent the combined work of many dozens of subject matter experts.

With PATs, the effort to curate existing knowledge stems from the belief that much of the technical work to achieve success exists – but expertise is fragmented, and repeated success is elusive. There is a perception that much development has been accomplished and there are now more sound building blocks from which the industry can implement new and repeatable work processes. This is in contrast to the perception nearly two decades ago that the CPTR was seen very much as a plan for research. As such, the PATs can be seen as a modern reinterpretation of the CPTR with more focus on business results and less on a highly technical vision for integrated and automated technologies.

3 Productivity Advancement Targets

As discussed above, Productivity Advancement Targets have been the principal work for Fiatech for the past few years. As with the development of the CPTR, there were initial workshops to develop the scope of the PATs and then workshops and meetings to develop each PAT. A more focused effort than the CPTR, the PAT teams were focused on productivity improving outcomes and had a motivating case study from an owner organization from which to build. This section reviews that case study, the PAT teams and sample enablers, and project recommendations resulting from the PATs.

3.1 Motivating Case Study

A large petrochemical owner-operator Fiatech member conducted an internal study and identified the potential to save $334 million on a $1.3 billion capital project – or about 25% of the total installed cost. This study first investigated the lifecycle costs of a plant and determined cost breakdown as percentages (see Fig. 3). Putting numbers to the percentages, a one billion USD plant translates to lifetime operating costs of 8.4 billion USD and transition costs of 300 million USD. As time in operation is only projected, operating costs are estimated while other costs are historical. The one billion USD in capital costs equates to 10 million for planning, 90 million for design, 400 million in materials, and 450 million in construction costs.

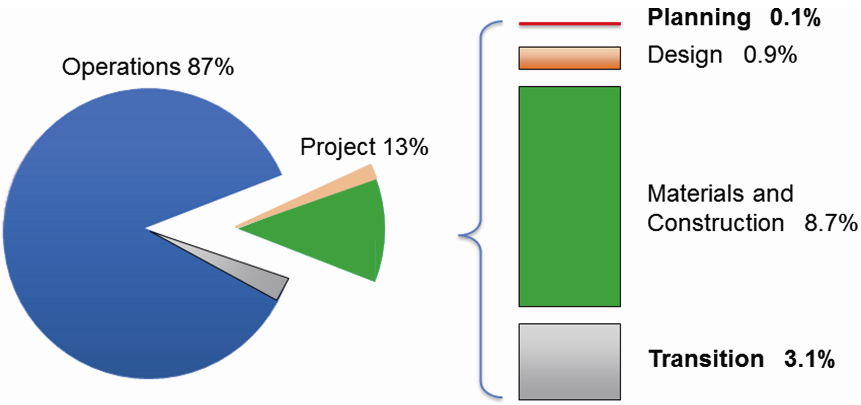
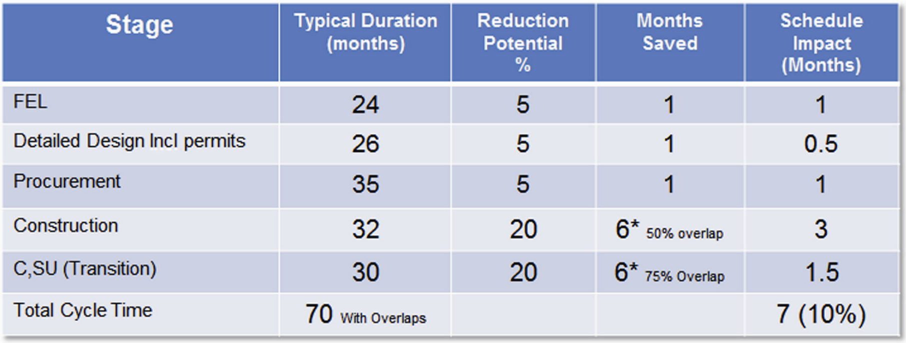


Fig. 3. Lifecycle cost breakdown of a large plant in the petrochemical sector, with total lifecycle costs in percentages of the total. (from [36])

As noted by the first author [37], principal savings come in reducing 300 million USD in transition (commissioning, startup, and information handover to operations) by 50% or 150 million USD. The substantial savings in transition costs stems from the repetition of data entry and poor handover of documentation. Complete electronic handover and proper setup for handover during planning should save considerable cost and time in the transition process. Another $149 million in savings is identified from $850 million in construction and materials costs with the balance stemming from efficiencies and better decision making to reduce owner-initiated change orders.

A related goal is to reduce total project time from inception to operations by 10% through shortening a 70-month schedule by seven months. Figure 4 shows the breakdown of the project timeline by major activity or phase as well as the estimate for potential savings with improved, technology-enabled work processes. Note on the figure, FEL is an abbreviation for front-end loading, often called front-end planning. The figure shows estimated schedule reduction in each phase, reducing front-end loading, detailed design and procurement each by one month and construction and transition each by six months. With overlapping activities, the net savings is seven months overall.



Stage	Typical Duration (months)	Reduction Potential %	Months Saved	Schedule Impact (Months)
FEL	24	5	1	1
Detailed Design Incl permits	26	5	1	0.5
Procurement	35	5	1	1
Construction	32	20	6* 50% overlap	3
C,SU (Transition)	30	20	6* 75% Overlap	1.5
Total Cycle Time	70 With Overlaps			7 (10%)

Fig. 4. Breakdown of estimated schedule savings in main project phase on a one billion USD capital plant, where phases overlap. (from [36])

While the goals of cost and schedule improvement are lofty, this case analysis is viewed by Fiatech members as believable, broadly generalizable to capital projects, and achievable given examples of realized technology deployments. The owner in question is using the analysis to drive improvement across its project execution processes.

3.2 PAT Focus Areas and Enablers

Broadly inspired by the case study above, Productivity Advancement Target teams were formed around thematic areas for improvement. There is not a direct mapping of PAT focus areas from the case study; instead, areas were generated by subject matter experts familiar with different aspects of capital projects and technologies to determine what would be feasible in the near or intermediate term. The overall goals of cost and schedule savings and productivity improvement from the case study were broken into smaller, more targeted goals. PATs are listed below with their associated goal:

1. Change Readiness. Improve sustained innovation and adoption by 55%.
2. Process Improvement/Reduction in Owner Change Orders. Achieve a 70% reduction in owner-initiated change orders.
3. Interactive Project Planning. Capture 25% improvement in project cost and schedule predictability (combined with PAT 4).
4. Integrated Project Management. Capture 25% improvement in project cost and schedule predictability (combined with PAT 3).
5. Regulatory Streamlining. Reduce review cycle times by at least 80%.
6. Design process improvements. Boost construction productivity by 10% and support improvement in handover and facility operations.
7. Integrated Materials Management. Improve construction productivity and reduce schedule delays by 20%.

8. Integrated Advanced Work Packaging and Information Mapping. Deliver 33% reduction in schedule and cost.
9. Automated Field Data Collection and Control. Enable 40% improvement in efficiency of field data collection and information access.
10. Interoperability. Increase the probability of productivity improvement success by 25%.
11. Facility Handover and Asset Management. Reduce transitions cost and schedule by 50%.
12. Emerging Processes and Technologies (aka Horizon 360). Accelerate identification, visibility, and industry development of technology by 40%.

PATs are meant to be synergistic. Achieving the goal of one is facilitated by the progress of the others. It is important to note that unlike the elements of the CPTR, there are explicit goals for individual PATs and capital projects overall. The realization of PATs is seen to be closer than that of the roadmap vision, due, in part, from the deployment of technology over the past decades.

As such, each PAT team was charged to curate existing knowledge and develop a set of enablers that can help organizations assess their capabilities and better plan investments. Each PAT team has lists of twenty or more enablers – in some cases many more. Enablers were broken into categories people, process, technology, continuous improvement, and contracting. Enablers were generally phrased to complete the question "Does your organization include." Sample enablers from PAT 7, Integrated Materials Management, include:

- (People) have sufficient staffing levels to support effective materials management?
- (Process) have a process for prioritization of vendor data collection and routing?
- (Technology) have technology in place to verify the "birth certificate" of materials to prevent the use of counterfeit materials?
- (Continuous Improvement) have a defined set of KPIs for materials management?
- (Contracts) have a contracting strategy that supports the approach to optimizing the overall system as opposed to individual silos?

The sample enablers selected provide a representation of the range of enablers for each PAT team. Enablers can be broad or very specific. Typically specific enablers represent more advanced capabilities; broader enablers are more likely to be present in most organizations.

3.3 Project Recommendations

While enablers represent recommendations to support the implementation of technology and processes to accomplish PAT goals, PAT teams also identified gaps in current capabilities. These gaps were used to recommend research and development projects. Each PAT team identified potential topics and refined them in a list of projects. These projects were then vetted at the Fall 2017 Fiatech Leadership Forum (also known as the Member's Meeting). The collected members received presentations from each PAT team on their observations and projects. The members voted for projects to prioritize them. From the meeting, the top ten projects were:

1. Real-time reporting for Advanced Work Packaging
2. Guide for Virtual Reality Construction Planning in Design
3. Guide to Adoption and Implementation of AWP and IAP
4. Data Specification for Consolidated View of Materials
5. Guide to Information Required in 3D Modeling
6. Reduce Change Orders Through Visualization
7. Integrated Materials Management & KPI Benchmarking
8. Critical Metrics for Strategic Technology Investment
9. Critical Attributes of Change Agents to Facilitate Tech
10. Handover to Operations Maturity Assessment Matrix

The list of projects is a representative sampling of those not prioritized. It is instructive that a number of projects are not technical in nature but instead focus on facilitating adoption through guidelines, metrics, and people. The sample is perhaps representative of the near-term charge of the PAT teams, but perhaps also is reflective of the belief that much technical work has been accomplished and the principal challenges are in implementation.

It is perhaps instructive to review the CPTR supply chain element projects (Fig. 2) with those listed above from the PATs. The PAT prioritized projects include four that touch on the supply chain – 1, 3, 4, and 7. Advanced Work Packaging (AWP) includes planning for supply chain through procurement work packages. Projects four and seven that discuss a consolidated specification for materials and materials management benchmarks are solely focused on the supply chain. These have direct analogs with CPTR projects shown in Fig. 2. This demonstrates that the opportunity for better engagement with the supply chain recognized at the founding of Fiatech has not been met. As noted in the case study, 400 million USD of the one billion USD capital spend is for materials; another 450 million USD in construction costs is directly influenced by the performance of supply chain.

4 The Future

With a focus on the PATs that more explicitly acknowledges the role of work processes, Fiatech members realized an overlap of interest with CII members and strength in research. For its part, CII members in the past two years showed renewed awareness of the importance of technology-enabled solutions. This serendipitous intersection of interests caused talks to occur about joint investment in research. Based on these conversations among Board members of both organizations, a merger was recommended to align the efforts of both groups fully. With near unanimity, membership of both organizations voted to merge in October 2017. Fiatech merged with CII effective January 1, 2018. The resulting organization will operate as CII, and the Fiatech name is retained as a specific focus.

The combined organization is a matrix of standing committees and sectors. Figure 5 shows the core aspects of the organization chart (administration and leadership committees are not shown). Supported by member volunteers and approximately 25 professional staff, the organization is centered around sectors (vertical bars in Fig. 5) Sectors represent industry affiliations – these are (1) Downstream and Chemicals, (2) Upstream, Midstream and Mining, (3) Power, Utilities, and Infrastructure, (4) Manufacturing and Life Sciences, and (5) Facilities and Healthcare. Fiatech is added as a sixth sector with the goal to focus on innovative projects. Sectors have their own funding in the organization and a great deal of autonomy to charter their own research projects. Each sector is supported by standing committees with specific areas of expertise, including Funded Studies, Implementation, Performance Assessment, and Professional Development. A Technology Awareness and Development committee is added to represent the practical technology expertise of Fiatech members and make that expertise available to Sectors. Fiatech PAT teams and CII Communities of Practice (COP) are combined into several Communities for Business Advancement (CBA) with subject matter experts in specific areas.

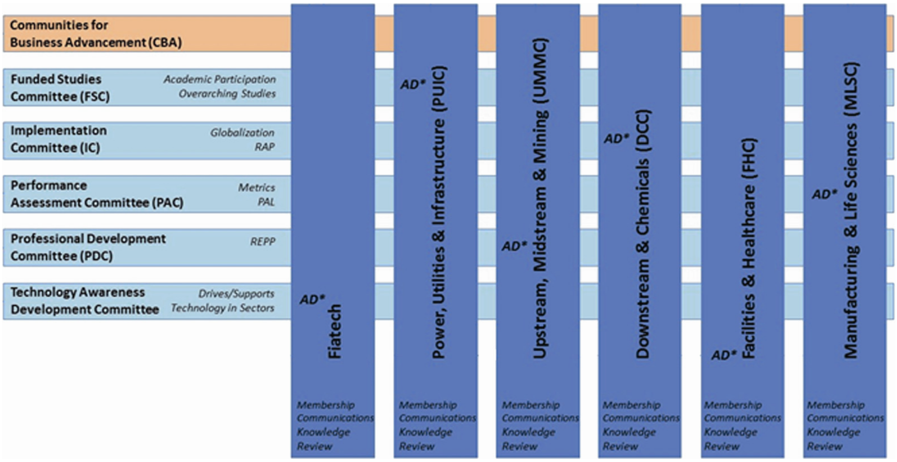

Communities for Business Advancement (CBA)
Funded Studies Committee (FSC)
Academic Participation
Overarching Studies
Implementation Committee (IC)
Globalization
RAP
Performance Assessment Committee (PAC)
Metrics
PAL
Professional Development Committee (PDC)
REPP
Technology Awareness Development Committee
Drives/Supports Technology in Sectors
AD*
Fiatech
Membership
Communications
Knowledge
Review
AD*
Power, Utilities & Infrastructure (PUIC)
Membership
Communications
Knowledge
Review
AD*
Upstream, Midstream & Mining (UMMC)
Membership
Communications
Knowledge
Review
AD*
Downstream & Chemicals (DCC)
Membership
Communications
Knowledge
Review
AD*
Facilities & Healthcare (FHC)
Membership
Communications
Knowledge
Review
AD*
Manufacturing & Life Sciences (MLSC)
Membership
Communications
Knowledge
Review

Fig. 5. Vertical Sectors and horizontal Standing Committees that comprise the principal elements of the combined Fiatech and CII organization. (from [38])

The combined organization will carry on the work of Fiatech; the projects emerging from the PAT teams will be brought into the Fiatech Sector and prioritized for funding. The CBAs will carry on the work of the PATs, and the Technology Awareness and Development committee will seek to advance near tear adoption of existing technologies. The plan is that the integration of Fiatech with CII will accelerate the adoption of new technologies and associated work processes. In turn, the original vision of the CPTR will be fulfilled, and the industry will see dramatic gains in performance.

5 Concluding Remarks

Fiatech has been an influential research and development consortium through its existence, despite never attracting the investment initially planned in the Capital Projects Technology Roadmap. As a consortium, Fiatech did help the industry move forward in information integration and interoperability, evaluation and adoption of now common field technologies, and regulatory streamlining. Perhaps most important, the CPTR roadmap vision as depicted in Fig. 1. is now something that is believed in by industry. At the founding, the vision was perhaps shared by academic researchers and technology evangelists but was not widely held by practitioners. The PAT areas, as a modern reinterpretation of the CPTR, represent a focused approach to technology and work process improvement that incorporates the developments of the past decades. Here, there are lessons for academic research. Perhaps most understudied is the area of information handover to operations. The case study identifying 50% savings potential should be a strong motivating example for further attention to this topic. Similarly, the apparent lack of meaningful progress to integrate the supply chain demonstrates the opportunity for substantive research. Finally, the implementation related gaps and projects – indeed, the rationale for the merger of Fiatech and CII – demonstrates the opportunities for a range of research on work process and human-centered approaches.

Acknowledgments. The authors acknowledge the contributions of the Fiatech members that have participated in various activities over its existence. Their effort made possible Fiatech and this brief history.

References

1. Hinze, J.: Making Zero Injuries a Reality. Construction Industry Institute, Austin, TX (2002)
2. RS 268-1: Front End Planning Tool: PDRI for Infrastructure Projects. Construction Industry Institute, Austin, TX (2011)
3. RS8-3: CAD/CAE in the Construction Industry. Construction Industry Institute, Austin, TX (1989)
4. SD-50: Cost Effectiveness of Computerization in Design and Construction. Construction Industry Institute, Austin, TX (1989)
5. Griffis, F., Sturts, C.: Three-Dimensional Computer Models and the Fully Integrated and Automated Project Process for the Management of Construction. Construction Industry Institute, Austin, TX (2000)
6. Fiatech History. http://fiatech.org/about/history. Accessed 12 Jan 2018
7. Fiatech: Capital Projects Technology Roadmap Initiative: Consolidated Roadmap, Fiatech, Austin, 207 pages, 30 December (2002)
8. Fiatech Capital Projects Technology Roadmap. http://fiatech.org/tech-roadmap. Accessed 10 Jan 2018
9. International Organization for Standardization. https://www.iso.org/standard/29557.html. Accessed 10 Jan 2018
10. Joint Operational Reference Data Project: Enhancing the PCA Reference Data Service (RDS) Operation. PCA and Fiatech (2014)

11. Dahl, T., Frees, S., Bickford, S.: Harmonization of Pump Schemas with the ISO 15926 Reference Data Library, Fiatech, Austin, TX (2011)
12. Fiatech – Capturing Equipment and Data Requirements Using ISO 15926 and Assessing Conformance (ERDC)
13. Koning, H., Willshaw, K., Temmen, H., Theissen, M., Teijgeler, H., Topping, R.: Mapping P&IDs: ISO 15926 Information Models and Proteus Mapping. Fiatech, Austin, TX (2017)
14. Akinci, B., Ergen, E., Haas, C., Caldas, C., Song, J., Wood, C., Wadephul, J.: Field Trials of RFID Technology for Tracking Fabricated Pipe. Fiatech, Austin, TX (2004)
15. Caldas, C., Haas, C., Torrent, D., Wood, C., Porter, R.: Field Trials of GPS Technology for Locating Fabricated Pipe in Laydown Yards. Fiatech, Austin, TX (2004)
16. Kang, J., Woods, P., Nam, J., Wood, C.: Field Tests of RFID Technology for Construction Tool Management. Fiatech, Austin, TX (2005)
17. Fernando, S., Bickford, S., Pawsey, N., Topping, R., Shah, V., Chopra, P.: Real-Time Field Reporting Using Smart Devices. Fiatech, Austin, TX (2012)
18. O'Brien, W., Mondragon, F., Howe, J., Sutton, T., Hijazi, F., Young, S.: Real-Time Field Reporting Phase II. Fiatech, Austin, TX (2017)
19. Song, J., Haas, C., Caldas, C., Ergen, E., Akinci, B.: Automating the task of tracking the delivery and receipt of fabricated pipe spools in industrial projects. Autom. Constr. **15**(2), 166–177 (2006)
20. Caldas, C., Torrent, D., Haas, C.: Using global positioning system to improve materials-locating processes on industrial projects. J. Constr. Eng. Manage. **132**(7), 741–749 (2006)
21. Mondragon Solis, F., Howe, J., O'Brien, W.: Integration of information technologies into field managers' activities: a cognitive perspective. J. Manage. Eng. **31**(1), 1–8 (2015)
22. Wood, C., McNeil, D.: Operations and Maintenance Information Needs Survey Results. Fiatech, Austin, TX (2003)
23. Kang, Y., O'Brien, W., Thomas, S., Chapman, R.: Impact of information technologies on performance: a cross study comparison. J. Constr. Eng. Manage. **134**(11), 852–863 (2008)
24. Kang, Y., O'Brien, W., Dai, J., Mulva, S., Thomas, S., Chapman, R., Butry, D.: Interaction effects of information technologies and best practices on construction project performance. J. Constr. Eng. Manage. **139**(4), 361–371 (2013)
25. Kang, Y., O'Brien, W., Mulva, S.: Value of IT: Indirect impact on construction project performance via best practices. Autom. Constr. **35**, 383–396 (2013)
26. Kang, Y., O'Brien, W., O'Connor, J.: Analysis of information integration benefit drivers and implementation hindrances. Autom. Constr. **22**, 277–289 (2011)
27. Phillips, T., Gould, B., Widney, J., Scott, B., Wible, R.: AutoCodes Project: Phase I, Proof-of-Concept. Fiatech, Austin, TX (2012)
28. Phillips, T., Gould, B., Widney, J., Scott, B., Wible, R.: AutoCodes Phase II Report. Fiatech, Austin, TX (2015)
29. Topping, R.: A Practical Deployment Strategy for Digital Signatures and Seals in Fully Electronic AEC Processes: An Overview of Digital Signature Technology and a Guide to Deployment. Fiatech, Austin, TX (2012)
30. Chasey, A., Pavelko, C., Root, S., Hogle, L.: Developing Core Technology Competencies: A Fiatech Research Report. Fiatech, Austin, TX (2009)
31. Fallon, K., Palmer, M.: Capital Facilities Information Guide Part I. National Institute of Standards and Technology IR7259 (2006)
32. Fallon, K., Palmer, M.: General Buildings Information Handover Guide: Principles, Methodology and Case Studies. National Institute of Standards and Technology IR 7417 (2007)

33. Demir, S., Garrett, J., Akinci, B., Bickford, S.: Requirements Recognition for Achieving Specification Automation. Fiatech, Austin, TX (2012)
34. Brawn, R., Nabavian, D.: Common Design Workflow Checklist. Fiatech, Austin, TX
35. Teague, T., Turton, R., Palmer, M.: Using XML Schemas for Facilities Equipment. Fiatech, Austin, TX (2004)
36. McNeil, D., Hunter, R.: Value Driven Approach to Productivity Enhancement. Fiatech Webinar (2016)
37. O'Brien, W.: Fiatech: the next generation of the capital projects technology roadmap. J. Constr. Eng. Manage. **143**(9) (2017)
38. CII Organization Chart.: https://www.construction-institute.org/CII/media/Documents/CII-Organization-Oct-2017.pdf. Accessed 10 Jan 2018

Advanced Construction Information Modeling: Technology Integration and Education

Raja R. A. Issa

Rinker School of Construction Management, University of Florida,
Gainesville, FL 32611, USA
raymond-issa@ufl.edu

Abstract. The Center for Advanced Construction Information Modeling (CACIM) conducts research related to all facets of advanced technology in the Architecture, Engineering, Construction and Operations (AECO) industry. Advanced construction information technologies comprises everything from Building Information Modeling (BIM) to Unmanned Aerial Systems (UASs) and reality computing technology. The study of such technologies includes not only the evaluation of the technology itself, but the study of workflows, use cases, and further development as well. The CACIM researchers work with industry professionals, regulatory bodies, and technology developers from around the world with the goal of discovering new ways that technology can be a part of the solution for improving the work quality and efficiency of AECO companies the world over in their quest to adapt to the 21st century. Current work of note includes the development of a Holographic Model Visualization System (HMVS); Virtual Reality workflows; BIM/VDC construction education development; Augmented Reality integration in construction; Reality computing processes and utilization; building code cost impact modeling and analysis; and strategic development of integration strategies for advanced construction technologies. Many of these projects have developed out of partnerships with local startup companies or agencies who see direct benefit from this work which also helps provide valuable insight to the CACIM researchers not only into the technology but also the associated workflows.

Keywords: Information modeling · Technology integration · Education

1 Introduction

The Center for Advanced Construction Information Modeling (CACIM), founded in 2010, conducts research related to all facets of advanced technology in the Architecture, Engineering, Construction and Operations (AECO) industry. Advanced construction information technologies, is the term used to generally describe the overall field of study and includes everything from Building Information Modeling (BIM) to Unmanned Aerial Systems (UASs) and reality computing technology. The study of such technologies includes not only the evaluation of the technology itself, but the study of workflows, use cases, and further development as well. The CACIM researchers work with industry professionals, regulatory bodies, and technology developers from around the

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 311–335, 2018.
https://doi.org/10.1007/978-3-319-91635-4_17

world with the goal of discovering new ways that technology can be a part of the solution for improving the work quality and efficiency of AECO companies the world over in their quest to adapt to the 21st century.

Over the past decade BIM and Virtual Design and Construction (VDC) have established an ever increasing presence in the AECO industry and many projects require the use of various advanced construction technologies. As the presence of technology continues to grow in the industry it is crucial that technology is continuously evaluated, improved, and developed in a way that encourages effective industry utilization. As such, the CACIM researchers focus on each technological development and subset of research with the goal of improving the AECO industry and AECO education as whole. To this end, CACIM is currently undertaking a wide range of research in partnership with industry, regulatory and development professionals in the following areas:

- Building Information Modeling
  - Education
  - Industry Application
- Reality Computing
  - Laser Scanning
  - Lidar
  - Infrared Scanning
  - Photogrammetry
- Augmented Reality
- Virtual Reality
- Holography
- Unmanned Aerial Systems (UAS)
  - Financial Feasibility
  - Industry Applications
- Corp. Tech. Integration Strategies

Each area of research represents work being done by individuals or teams of graduate students, at both the Master's and PhD levels, in collaboration with faculty and industry practitioners. Current work of note includes the development of a Holographic Model Visualization System (HMVS); Virtual Reality workflows; BIM/VDC construction education development; Augmented Reality integration in construction; Reality computing processes and utilization; building code cost impact modeling and analysis; and strategic material development of integration strategies for advanced construction technologies. Many of these projects have developed out of partnerships with local startup companies or agencies who see direct benefit from this work and help provide valuable insight not only for the work but for the faculty and students conducting it. The goal for every research project undertaken by CACIM is to make a meaningful impact or contribution to the body of knowledge, with the addition of improving upon current industry and academic processes. Each area of study is carefully selected to be at the forefront of technological innovation within the AEC industry. Over the years, CACIM researchers have contributed substantially to the body of knowledge through publications, advancement in AEC educational pedagogies, and advanced research into technological advancements within the industry.

2 CACIM Research

2.1 BIM in Graduate Education

The research conducted at CACIM spans a range of technological disciplines. As the AEC industry evolves and continues to become more technologically integrated the number of areas in need of study continues to grow. Currently, CACIM is engaged in research in each of the major areas of study which are leading the industry and require the greatest attention at this point in the industry's development. Furthermore, each area of study continues to evolve as the needs of the industry change. Work to further develop the technologies themselves is common, followed by research related to improved pedagogical techniques. This work is aimed at improving the education of the next generation of industry professionals, who need to be well versed in technologies which will impact their careers. In this way, CACIM strives to not only conduct research which is impactful and meaningful to the industry but also to ensure that such advancements are able to be adapted by the industry itself. While a great deal of work has emerged from CACIM since its inception, a sample of that work is used herein to demonstrate the industry leading quality of the research.

Construction technologies in all forms have found an increasing presence in the AEC industry as well as in AEC education. At the University of Florida this reaches its apex in the graduate level "Construction Information Systems" course. This course requires students to complete a series of BIM models and to familiarize themselves with a number of secondary and tertiary construction technologies [1]. Furthermore, the course focusses on the technological workflows and industry practices which are an important part of moving towards greater efficiency and better projects through the integration of technology. The process is the key and students are pressed to work in simulated real-world scenarios while completing modeling exercises which are for an actual client.

The course meets weekly for a single 3 h time block, over the course of an academic semester which typically runs a total of 15 weeks. It is an upper division course focused on the utilization and management of BIM and VDC technologies. Students are expected to have a basic understanding of construction practices and basic computer/technological proficiency as well as an adequate personal computer as per the College and School admission requirements. The software platforms primarily covered in the course include modeling software, drawing management software, 4D and 5D BIM platforms, as well as data management tools. During the course students complete seven homework assignments, based on standard plans designed specifically for the course, as well as one semester long group project developing an as-built model of local buildings near the University (see Fig. 1).

The primary focus for the individual assignments is on best practices, theory, and BIM execution, while the group project focusses on team collaboration and BIM workflows over the duration of a project. Essentially, the core components of skill building and concept application drive the two sets of activities which make up the backbone of the course. The collaborative exercise bridges the gap between skill building and application related to connected BIM services and processes.

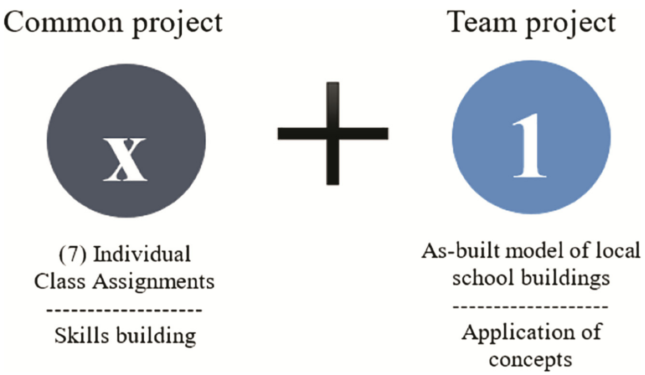


Fig. 1. Class work

The course begins with instruction and utilization of desktop applications and students are encouraged to work in the computer lab as well as on personal computers. The primary focus in the early stages of the course is on modeling best practices, information management, model management, and the development of foundational BIM skills. These skills are then built upon toward the goal of achieving literacy in connected BIM processes, which most effectively prepare students for their eventual career in the AEC industry. The move from desktop stand-alone application to connected BIM occurs in stages, each leveraging a different piece of technology. The stages of technological integration and exposure can be described on a scale of good, better, best, as they relate to the level of integration and exposure achieved in a course.

At the early stages the use of models to coordinate and update designs is something which can easily be added to the fundamental BIM skills developed by the students. Next would be getting the students to leverage document and model management technology throughout their assignments. Finally, the students achieve collaborative BIM literacy through the use of collaborative modeling practices, coupled with integrated document/model management, and coordination capabilities.

The most effective way to achieve connected BIM literacy is through the use of hands-on activities and exercises which afford the students the opportunity to put their technical skills to use and experience the workflows first hand. In that spirit, it is difficult to achieve meaningful examples of connected BIM utilization, with any cloud based platform, intended for disparate project teams, in a classroom setting. The development and execution of collaborative exercises with industry partners can be an impactful way to demonstrate the power of connected BIM while working through an engaging and meaningful learning experience for everyone involved (see Fig. 2) [2].

A collaborative BIM exercise was added to the course to provide students with a real-world BIM coordination experience which exposed them to the workflows they will be a part of during their career. This exercise was designed to provide students with a meaningful experience using a range of connected BIM technologies, while simultaneously affording the industry partner the opportunity to train new employees or test new workflows in a no-risk environment. The bilateral benefits made possible through this exercise can continue to evolve and grow as the academic and industry partnership

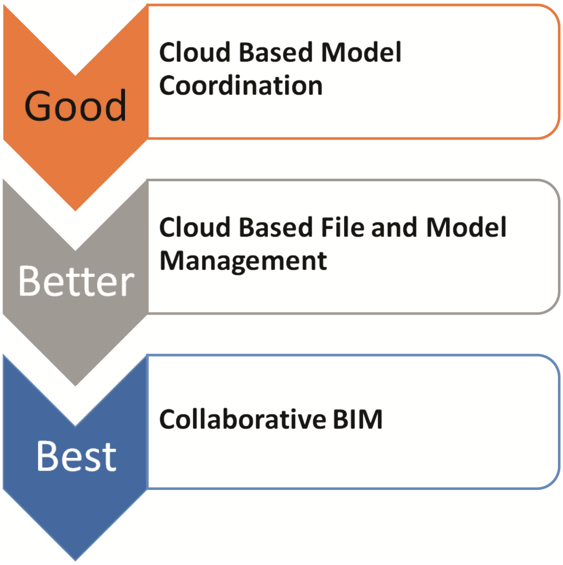

Good
Cloud Based Model Coordination
Better
Cloud Based File and Model Management
Best
Collaborative BIM

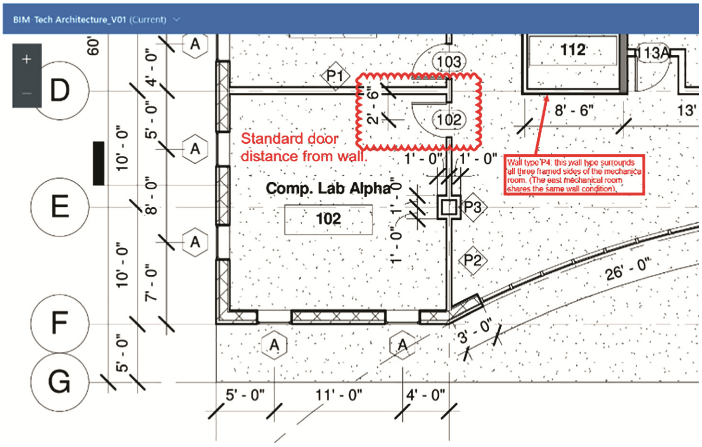
Fig. 2. Connected BIM

evolves and change is needed. Whether the person participating in this exercise has never experienced connected BIM enabled coordination or simply wants to expand their understanding and experience to new areas, the focus on experiential learning creates meaningful experiences which encourage further thought and skill development.

The benefits of an interactive, synchronous BIM based coordination activity cannot be overstated. Students gain a firsthand understanding of the complex process while developing a respect for the roles of each member of a project team. The ability to assume the role of a specific team member and see the problem from a new perspective, with a tangible goal in mind, forces the students to think critically and apply their knowledge of the BIM processes in a real-world scenario. The ability to utilize the technology, concepts and intangible skills, such as communication and negotiation, which will be crucial in their careers is a great asset to their education. Furthermore, it affords them the opportunity to make mistakes in a low-risk environment where instructors and peers can offer insight and guidance toward success. Overall, the ability to work synchronously in a connected BIM environment provides benefits which break down the traditional asynchronous stereotypes placed upon learning while enriching the students overall learning experience. Beyond in-class exercises CACIM strives to constantly improve students learning experiences and grow the curriculum as well. One additional way this has been achieved is through the integration of drawing management software in the classroom.

As BIM becomes further imbedded in the AECO industry workflows and as such an emerging area of concern is the transmittal of data present in the model to the field. Furthermore, the use of and navigation through 2D drawing sets, especially when developing as-built BIMs, can hinder the productivity of Virtual Design and Construction (VDC) processes. With increased expectations that college graduates in AECO fields achieve competency in VDC processes, instructors of BIM courses have the unique challenge of teaching both modeling techniques and best practices to students. While students are expected to interpret construction drawings sets and develop a completed BIM model inclusive of all building systems, it was determined that students were spending a great deal of time trying to organize and navigate disparate drawings sets (sometimes exceeding 750 sheets), rather than spending that time to develop their VDC expertise. Due to this, a drawing management platform was integrated into the course in order to aid students in their work; provide greater content control to the instructors; and to create exposure to the tools the students will be expected to use when they begin their careers.

To test the effectiveness of this approach, CACIM researchers conducted a survey to establish the approximate amount of time which students spend navigating drawings as part of the BIM development process from 2D plans. The class in the subsequent semester was provided access to an emerging drawing management software tool, PlanGrid [3]. This tool organizes drawing sets by detecting and displaying relevant sheet identity information, along with standard drawing set navigation markers, e.g. section cuts or callout boxes. It then intelligently links these navigation markers to the appropriate pages and allows seamless navigation between drawings on a web-based platform, accessible through any Internet ready device or PC (see Fig. 3). Students provided with access to the drawing management tool were asked to evaluate their experience and provide initial reactions to the use of the software. The utilization of advanced drawing management software in the BIM classroom allows students to focus more of their time on the development of BIM/VDC knowledge, enabling them to be better prepared for their careers in the AECO industry. Furthermore, it helps illustrate the importance of organized, clean model and data management.



Standard door distance from wall.

Wall type P4: this wall type surrounds all three framed sides of the mechanica room. (The east mechanical room shares the same wall condition).

Fig. 3. Markups in drawing management tool

In addition to drawing management the CACIM researchers have explored other techniques to improve educational experiences. One such improvement was made through the inclusion of keystroke capture recordings of the in-class lectures [4]. Instructors of BIM courses have the unique challenge of teaching software packages to students, both undergraduate and graduate, which challenge their preconceived notions of 2D plan development. Furthermore, students tend to have a hard time following along with the instructor while simultaneously taking notes to remember the steps required for various modeling activities. This can lead to missed steps and the need for the instructor to repeat steps multiple times to help each student master the skills being taught. Autodesk Screencast was used to record the computer screen and the user keystroke input in the BIM software, eliminating the need to write down the steps taken to accomplish various modeling activities (see Fig. 4). The goal of this research was to assess the impact of using screen capture software on the students' BIM learning experience.

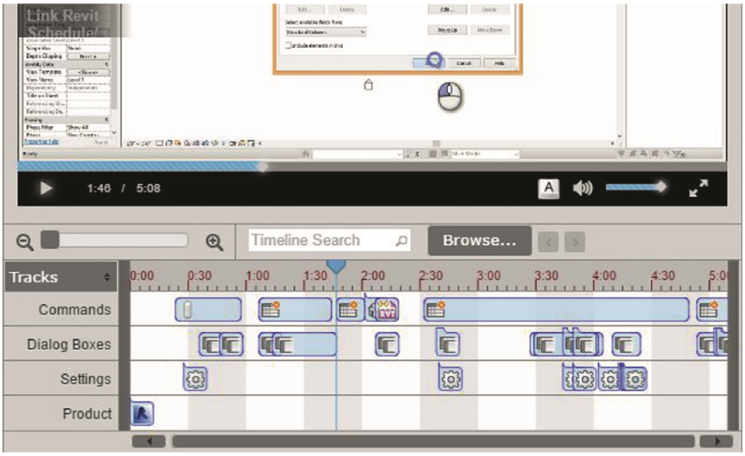


Fig. 4. Screen capture software playback interface

During the instruction of BIM software, students were introduced to the screen capture software and instructed to use it to record their work during class. The integration of screen capture with the BIM platforms, coupled with web-based storage, allowed students to record and access recorded data from any computer with Internet access that they used for their assignments. In addition to their individual recordings, the instructor station screen and keystrokes were also recorded during the lectures in order to allow students to go back and review the capture of portions of the topic which the students, in that particular class, were having difficulty with. These recordings were supplemented with recordings created by the instructor outside of class based on the student requests. This method departs from recording canned tutorials and focusses on the needs of the specific students with respect to each topic. Furthermore the students were able to use the recordings from the class as a reference when working on assignments and also for future courses. Interestingly, during the study students began requesting additional screen capture recordings. The recordings removed the need for students to try to take detailed notes on the "button-pushing" aspects of the lecture and allowed them to focus on following along and mastering the concepts and best practices being taught.

Initial indications for the use of this technology were positive, with the average usefulness rating by the students being an 8.2 out of 10 and the majority of students reporting that they used the recordings multiple times. Through a survey, the students indicated that the screen capture recordings were useful as they worked on assignments citing both the keystroke capturing and ability to review the recordings on their own time as the most beneficial aspects. Keystroke capturing provides a definitive account of exactly what shortcuts or tools were used in class during various modeling activities. These are the most common questions raised during a class and office hours on subsequent days. During the architectural and structural modeling portions of the course the instructor saw a decrease in technical software questions during office hours and students were instead asking about more advanced BIM concepts related to the imbedded information in the model elements. These are all positive signs that keystroke capture software may aid students' modeling abilities, allowing instructors to focus on the important conceptual and advanced topics in BIM education. Furthermore, the researchers found that as students continue in their education they often return in later semesters with questions related to BIM components which are integrated into their other courses. Thus the recordings can serve as a resource for the students throughout their academic and professional careers beyond the course itself. New methods and technologies for the instruction of BIM, such as those reviewed in this here are continuously sought after and researched CACIM. This is one important aspect of the work conducted at CACIM as educators look for the most effective methods for enhancing the educational experiences of their students to help them towards long term career success.

2.2 Augmented and Virtual Reality

Augmented Reality (AR) is the superimposition of computer generated digital information over real world views [5]. The study of AR has spanned many industries and has continued to evolve. The architecture, construction, engineering and operations (AECO) industry has begun to explore applications for AR and many researchers are completing work in this regard. Some of the areas in which researchers are deploying AR include: as-planned to as-built progress monitoring, training, dynamic site visualization, construction defect detection and integration with various building information modeling (BIM) workflows [5, 6].

CACIM researchers conducted a study, sponsored by the National Science Foundation (NSF) [7, 8], using Augmented Reality (AR) technology, reality augmented with a layer of artificial visualizations (virtuality), to simulate and enhance the project context and spatio-temporal constraints of masonry [7], steel structure erection [9], and roofing application [10]. The goal was to determine whether learners would better comprehend the elements and hidden processes which exist in complex construction sequences. The future of the construction industry is highly dependent on the competence of new employees. Therefore, it is crucial for new employees to enter the industry with the abilities required to comprehend intricate and complicated problems inherent in the construction process. However, Construction Management students often receive inadequate exposure to in-situ construction processes and procedures, which can be detrimental to their early success and ability to effectively solve problems. In this regard, students often lack a thorough understanding of the complex spatio-temporal constraints which exist during the construction process.

Through the integration of AR technology, the spatio-temporal constraints are enhanced thus enabling learners to visualize context and hidden processes. A significant improvement of the perception of reality can be expected through the combination of the learners' ability to understand the complexity of construction products (e.g., assemblies) and associated jobsite processes by using the real environment augmented with computer-generated information layers. The superimposition of images serves as an instructional mechanism to virtually incorporate jobsite experiences into the classroom. In this study, 3D generated models (virtuality) were superimposed on videos (reality) documenting a construction project in each of the steel structure, exterior wall and roofing phases (see Fig. 5). The video with the augmentation was then used to test the effect of such technology on students' comprehension of the elements in these building assemblies. This study was also extrapolated to the use of AR technology for training and comprehension within the AEC industry itself.

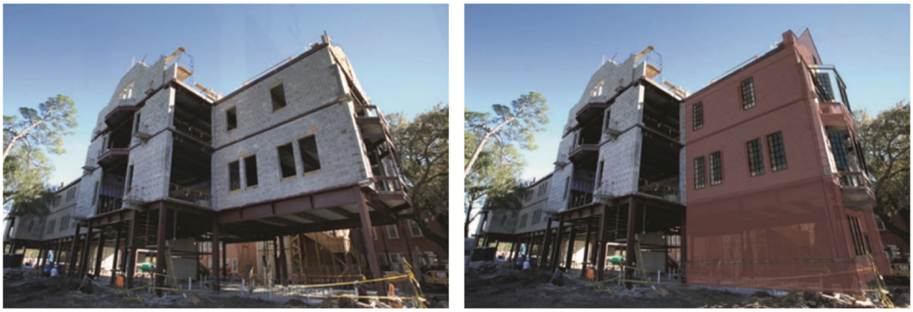


Fig. 5. Completed BIM model augmentation over on-site construction progress documentation

Testing demonstrated that students exposed to AR enabled content were better able to identify the elements and tasks related to steel, masonry and roofing construction when compared to those students who received only standard in-class instruction. Furthermore, the best outcome was achieved by exposing the students to the augmentation video as a supplement to the class instead of a replacement.

Moreover, the level of augmentation affected the level of understanding and retention for students. The results indicated that the students' ability to recognize the "brick ties" element was impacted more by the augmentation than their recognition of the "flashing" element. That can be attributed to the fact that the brick ties had higher level of augmentation in the video whereas the flashing was only highlighted in limited fashion.

The implementation of augmented reality (AR) technology for construction management applications is rapidly evolving and being considered for use in many new applications. This research conducted by CACIM was a first step in determining the effectiveness of AR for enhancing construction education experiences and discovering how to optimize its use moving forward. The ability to further develop students' spatio-temporal understanding of the complex problems which pervade the construction industry is a crucial part of preparing them to enter the workforce. Furthermore, training within the industry can be delivered in the same way and reap the same benefits of AR integration as students in the classroom. AR has shown positive potential in enhancing construction management education and training within the AECO industry. During each of the masonry, structural steel and roofing portions of this study it was discovered that AR was able to help students better understand and identify some detail elements involved in those assemblies. Furthermore, it was found that there was an advantage to combing traditional classroom lectures with AR enabled media. As AR technology continues to evolve and become interconnected with virtual reality (VR), the potential use cases and impact it can have on the industry will grow exponentially.

2.3 Holography

In the AECO industry, different visualization techniques are used to represent a project and novel visualization techniques continue to emerge. These techniques range from printed two-dimensional (2D) drawings to full 3D BIM displays. Recently, the AECO industry has seen an increase in the research and application of wearable VR and AR technology for project visualization as well. CACIM researchers developed and evaluated a new visualization technique, a Holographic Model Visualization System (HMVS), which does not require wearable or individual devices, enabling group collaborative visualization (see Fig. 6) [11]. The developed technique uses a less-expensive technology based on creating holographic illusions directly from BIM models. The hardware is designed to be used on a meeting table where the hologram can be seen from any point and manipulated by meeting attendees using hand gestures and/or voice commands. Content for the proposed device is developed directly from BIM models and imported into the visualization environment. The developed prototype includes sample projects where the user can rotate the building by floor, as well as compare models of each construction discipline, thus enabling model coordination and review activities. The ultimate aim of this research is to develop the HMVS to seamlessly interface with existing BIM models, to facilitate the understanding of project-related issues in on-going projects or in academic settings. Additionally, researchers are working to develop add-ins and workflows which enable the direct export of a BIM model for display directly via the holographic projector, without the need for additional user manipulation. Moving forward this research will be expanded to include the testing of the developed HMVS's performance, as well as the ability to detect clashes, change materials, and compare planned versus actual construction progress.

Effective communication of information is an important factor for the execution of a successful project. The developed HMVS displays BIM models with functionality outside of the native modelling environment by producing a project visualization that is user-friendly and functional in group settings. The projected model can be manipulated using voice commands and hand gestures in order visualize all aspects of the model, detect major clashes, filter for specific building elements, or to visualize complex building sectors. Both the hardware and the software of the developed technology are undergoing further development toward the inclusion of additional functionality and to improve its efficacy.

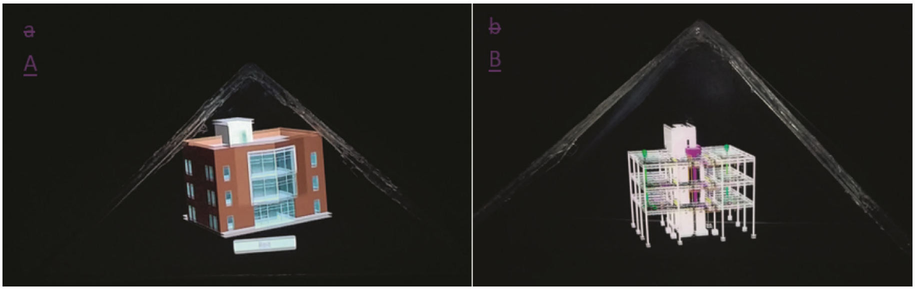


Fig. 6. BIM models displayed via holographic projection

The HMVS is being considered for a wide range of applications including as a presentation tool, for coordination meetings, or as a learning tool within AECO academic programs. As BIM continues to become more prevalent in the AECO industry, it is crucial to develop ways to improve access to the data, both visual and informational, which is inherently built into BIM models to enable more efficient data driven decision making throughout the lifecycle of a construction project. This research has developed a novel system which provides a means of accessing and displaying the information rich content in a BIM model in a user-friendly environment free of the confines of the BIM development environment.

2.4 Reality Computing

Reality computing is an area of research and development within the AECO industry which has grown substantially in recent years. Reality computing at its core is the process of bringing real-world 3D information into the digital world for evaluation and use. This takes on many forms including; laser scanning, photogrammetry, light detection and ranging (Lidar), as well as UAS enabled data capture. These technologies enable real world conditions to be captured, with varying degrees of accuracy, speed and cost, and used as part of increasingly technologically enabled construction projects. The applications for such technology extend beyond construction and into historic preservation, landscape management, and site controls as well. A multitude of industries are finding value in the ability to capture real-world conditions and merge them with digital models, however the construction industry stands to reap the most benefits from such technology. This is not lost on the CACIM researchers who have explored reality capture for educational training, technology evaluation, and workflow improvement.

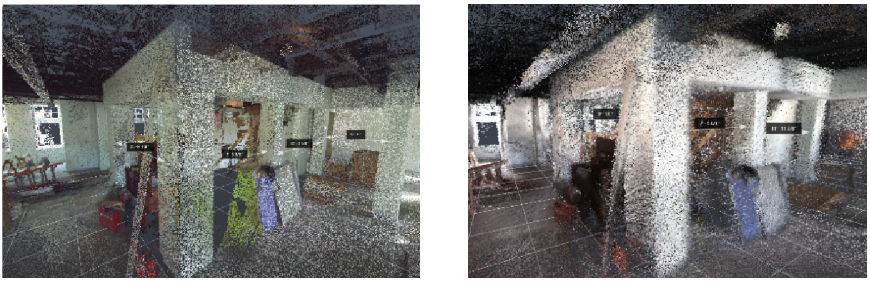
Laser Scanning and BIM. Laser scanning technology, a popular reality computing technology, has emerged as a useful tool in documenting existing conditions of buildings. The main application for such documentation is to assess current as-built conditions of existing, mostly historical, buildings. The technology can also be used as an integral part of construction progress documentation in new projects. In order to equip students with knowledge on the latest technologies in the industry, laser scanning was introduced to students in a BIM class [12]. Students were given a thorough demonstration on how the equipment functions. After that, several campus buildings' exteriors were scanned. Students had the chance to see the scanning in action. More importantly, the generated point-clouds from the scanner were given to students as part of the final project. Five teams were formed and each team was given a set of fire-evacuation floor plans, with minimal details, of an existing campus building. The teams were then asked to take measurements from the point-clouds and model the buildings in Autodesk Revit. The lack of marked elevation drawings did not prevent the teams from modeling the buildings fairly accurately. The point-clouds gave the students an large amount of information that were lacking in the provided documents. Even though the learning curve for navigating such point-clouds was a bit steep, the extracted information helped the students better understand the buildings' envelopes and realize the real value of laser scanning technology.

The projects assigned in the course would have been impossible to model as accurately without the provided point-clouds. The lack of elevation and section views presented challenging obstacles, especially when trying to determine the floor levels. Taking actual site measurements were also very difficult, particularly for the exterior façade that in some cases extended four stories high. Therefore, students experienced the real value of the generated point-clouds. Having hands-on experience on getting the required information to model a building from a point-cloud gave students a new perspective and appreciation of the technology. Exposing students in the construction management program to the latest technologies in the field gave them a competitive advantage in the job market. In addition, the industry would pick on these new technologies much faster with graduates that are familiar with their operations and uses. That would also lead to more enhancements and innovations to make these new technologies even more efficient. The study showed that students are eager to try new methods and approaches in the classroom as they come to market.

Scanning Technology Comparisons. As reality computing technology becomes increasingly popular throughout the AECO industry, novel technologies and solutions are emerging to provide alternatives to traditional workflows. Traditional phase-shift or time of flight laser scanning devices have been the standard in the industry and set the bar for data accuracy. However, these devices can be costly and time consuming to operate, which limits their availability for many projects. To counteract this concern, there has been a surge in emerging technologies that use infrared, Lidar, photogrammetry or a combination of those solutions [13]. The workflows of these emerging scanning techniques are vastly different from that of traditional laser scanning, which is the primary advantage these solutions rely on. Emerging point cloud generating technologies show great promise for the AECO industry, but need to be effectively evaluated to achieve widespread acceptance. The goal of this research was to assess one of these emerging reality capture technology and compare it to traditional laser scanning. A mobile Lidar based technology was evaluated and compared to traditional laser scanning based on a number of factors including; accuracy, operation time, processing time, workflow and quality. A field study was conducted using both devices and associated processes. The data collected were used to compare and evaluate their performance.

This research provides a baseline understanding for the comparison of a mobile Lidar scanning solution and traditional laser scanning techniques in the AECO industry.

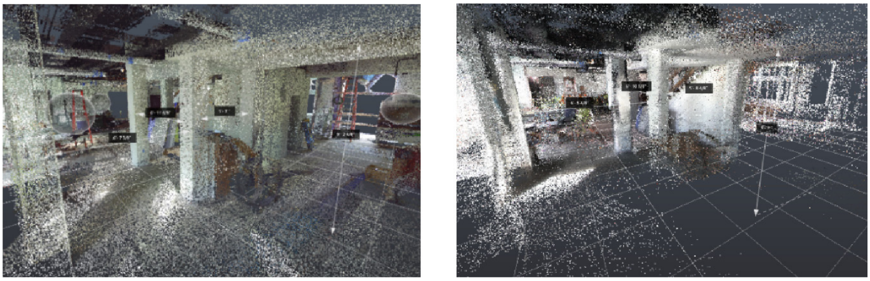
The comparisons made in this research offer many factors for consideration when evaluating reality capture devices for potential use. Traditional laser scanning and mobile Lidar systems demonstrated the largest variance when it came to workflow and time. The mobile Lidar system demonstrated a 65.1% reduction in time, which could have a great cost implication on a jobsite. Additionally, the workflow offered the advantage of enabling the operator to walk the target area while capturing one continuous data stream. The less time consuming and more fluid process of the mobile Lidar system may be able to fit better into existing workflows. That being said, the technology must undergo further refinement in order to improve its accuracy and usability on a jobsite. The accuracy of the mobile Lidar system used was, on average, 1 21/64” less accurate compared to traditional laser scanning. This must be taken into consideration and could be excessive depending on the intended use for the data. Figures 7 and 8 show samples of the data output from each of the systems being compared. Traditional laser scanners can be cost prohibitive as they require a significant investment up front. New laser scanning devices which bring this cost down into the $15,000 to $25,000 range. The mobile Lidar scanner used in this study currently markets for $25,000, putting it in the same price range as the newly unveiled Leica and Faro traditional laser scanning devices.



a. Traditional Laser Scan Point Cloud

b. Mobile Lidar Based Point Cloud

Fig. 7. Main staircase columns



a. Traditional Laser Scan Point Cloud

b. Mobile Lidar Based Point Cloud

Fig. 8. Point cloud of building south column array and ceiling

There is a wide range of uses for reality capture data and each factor compared in this study should be taken into consideration when determining the potential use cases for a device. A great deal of research is being conducted to define further use cases for reality capture data and such research will parallel the refinement and development of new technologies. Technology, such as mobile Lidar scanning, are expanding the world of reality computing in the AECO industries. The desire for better access to data and means of integrating the virtual and real worlds continues to grow as technology evolves. Mobile Lidar technology will continue to be refined, and as such project teams may find it feasible to compare construction progress on a daily basis by generating a daily, as-built model of their job which can be compared to the as-planned BIM. Traditional laser scanning revolutionized the way reality capture took place and will always have an important role in the development of highly accurate and detailed as-built models. However, the CACIM research findings suggest that mobile Lidar technology may offer an efficient alternative and supplement which could expand the capabilities of project teams throughout the AECO industry.

Planning Laser Scanning Work. The use of point clouds generated by laser scanning, one technology within the space of reality computing, allows for accurate as-built documentation, construction verification and analysis. However, the use of the data can be hindered by poor scanning techniques or a lack of planning prior to the start of a scanning job. The proper development of a laser scanning plan guides the operator through setting up and locating appropriate scan locations for a given building or space. Laser scanning experts often develop such plans quickly by walking around the site, as they are well versed in the limitations and requirements which influence the scan outcome. Professionals with little or no scanning experience rely on the manufacturers' recommendations, advice from experienced scanning professionals or trial and error to develop a thorough understanding of how to best execute a laser scanning job. This process can lead to costly re-scanning and time consuming troubleshooting. CACIM researchers have recognized this problem and ventured to develop a simple tool to help alleviate some of these potential errors. Through this research a parametric scan planning tool (PSPT) was developed to aid in the visualization and development of a scan plan for laser scanning jobs [14]. PSPT has been designed to accommodate a range of scanner capabilities through customizable properties. It also encourages 3D visualization of scan boundaries within the building or space being scanned.

Planning work is an important part of the lives of professionals in the AECO industry. A typical surveying crew, for example, will spend time planning a job before ever stepping in the field. Why should laser scanning be any different, being that it is a means of capturing data out in the field similar to some surveying operations? Figure 6 shows a rudimentary scan plan that was developed to scan a fairly complex building. This kind of plan is typical of what is done, but it is missing a few key things. First it has no scale, meaning that the locations of the scans are based on the professional's knowledge of the building's size. Secondly, there is no indication of the scans "sphere of capture" or the effective radius of each scan. In this regard, the plan is a rough estimate of what may take place but does not visually or numerically indicate what each scan location is intended to capture. This plan is a great place to start and can be completed in the field if needed. However, any mistakes made in the planning of a scan can lead to the incorrect estimation of the quantity of scans needed and the time required to complete the project. If the plan is incorrect and field adjustments are not made there can also be an incomplete data set, leading to point cloud registration errors due to gaps in the data. None of these outcomes are ideal, as the ownership and operation of a laser scanner is expensive and it is in everyone's best interest to be as efficient and effective as possible.

The scan plan shown in Fig. 9 was re-created in Fig. 10 using the parametric scan planning tool (PSPT) developed through this research. Figure 10 shows the completed scan plan with capture radii of 30 ft. and 60 ft. respectively. Upon completion of the plan using scanner settings with a 30 ft. effective scan radius it became clear that the scanner locations selected were inadequate. There was virtually no overlap and the scans would be ineffective. Given this new information it was determined that there were two solutions; (1) the settings of the scanner could be increased to provide a larger effective scan radius or (2) the scan locations could be re-evaluated and an increased number of scans could be conducted. Figures 7a and b show the same scan plan accounting for the adjustment in settings which lead to an increase in the effective scan radius from 30 ft. to 60 ft. The 60-ft. radius provides adequate coverage and overlap for the majority of the building and a few additional scans may be necessary to fill in the remaining gaps.

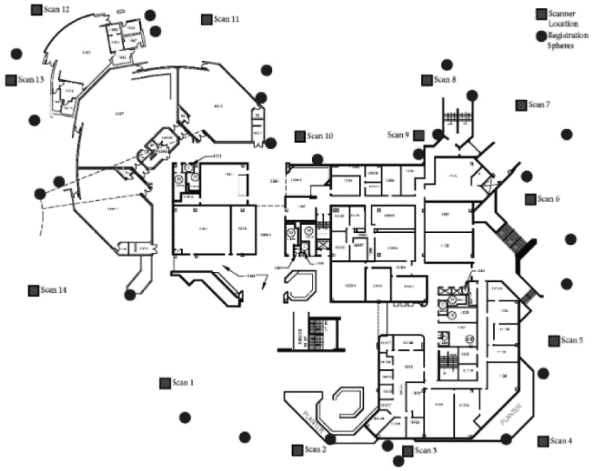
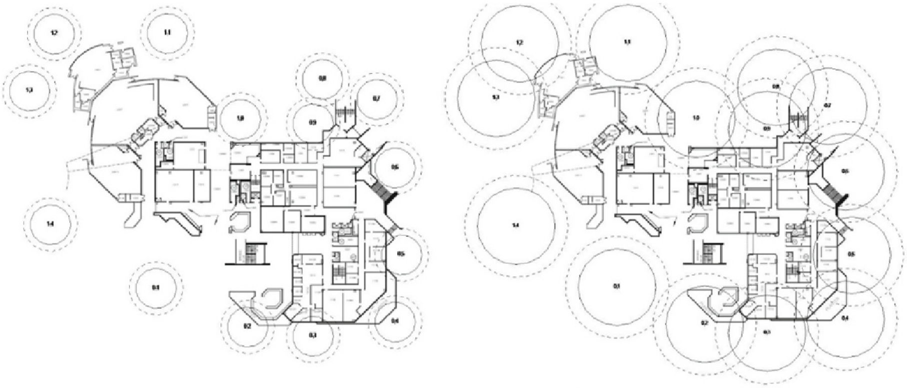


Fig. 9. Rudimentary scan plan, with scan and target sphere locations

Figures 10a and b demonstrate how the developed PSPT was able to help inform planning decisions and aid in the development of a realistic scan plan. Every decision made had a time consideration that impacted the total duration of the scan operation. Increasing the scanner image resolution quality settings led to an increase in the individual scan time and adding scans added to the cumulative scan duration. All of these factors can play a part in determining if and when scanning operations are to be conducted and can help the parties involved realistically estimate the expense and time implications involved. The PSPT was used in planning for the example outlined in Fig. 7, which would be similar to an as-built documentation study where a model does



(a)Capture radius of 30 feet

(b) Capture radius of 60 feet

Fig. 10. Scan plan using the PSPT

not currently exist, however it can also be effective in 3D to assess the amount of vertical building which will be captured during the scans.

The 3D visualization made possible through the use of the PSPT adds a new layer of information to plan view utilization, showing omissions in the data which cannot be seen in plan views. Figure 11 shows the plan view of a building where the PSPT has been used to plan for the scanning of the main façade of the building. The entire façade can be captured in five scans with more than adequate overlap. However Fig. 12 shows the proposed plan in 3D where it is visible that the selected scanning settings and locations will only capture roughly 75% of the building façade. Given this information it is possible to make adjustments to the scan settings to ensure that the entire façade is captured and to verify that the developed plan will yield the desired results.

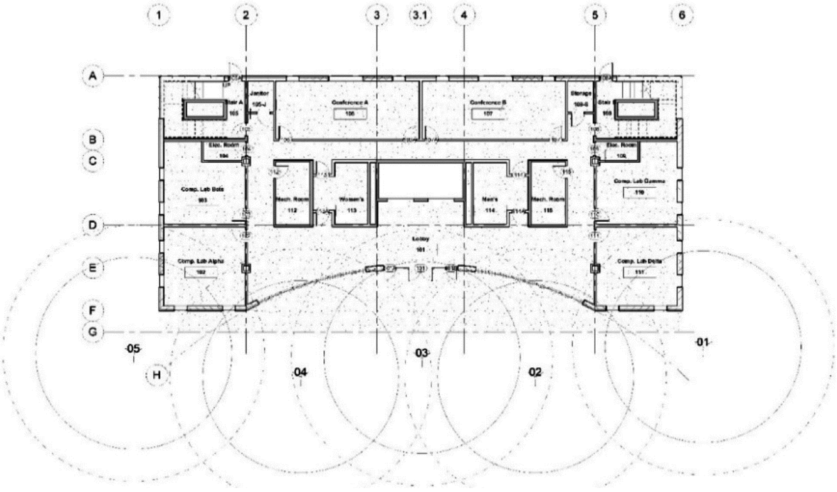


Fig. 11. Plan view of the completed scan plan for a building façade using PSPT

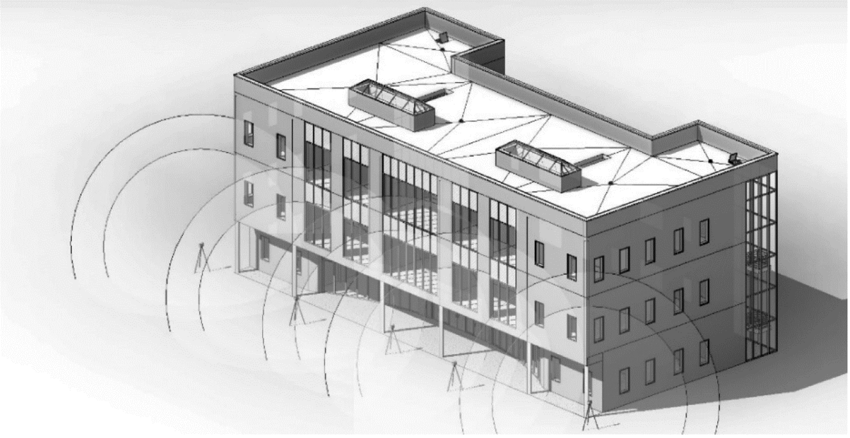


Fig. 12. High angle 3D view of the completed scan plan for the building façade using PSPT

Planning for a laser scanning job is a multi-faceted endeavor which, when done correctly, can help avoid frustration when the scans are not adequate and a second visit to the site is necessary for re-scanning. The PSPT aids in identifying the locations for scan set ups but can also help inform decisions regarding the desired settings which should be used in the field. The PSPT can not only help AECO professionals but also students as they learn new skills and develop the expertise to plan laser scanning jobs. PSPT can help reduce the learning curve that users often face when first starting with laser scanning. It provides a visual cue and a scaled representation of the scanners capabilities which can be difficult to grasp for those beginning their work with laser scanning. While additional work is required to refine the PSPT and tailor it to the needs of the various parties within the AECO industry, this research has provided a crucial first step in the development of an effective planning tool for laser scanning and reality computing operations.

2.5 Unmanned Aerial Systems (UAS)

Unmanned Aircraft Systems (UASs) are an emerging technology with a variety of applications, which extend beyond the realms of the military and hobbyist markets. Many industries are trying to decide if UASs can be utilized to complete tasks in a safer or more efficient manner. In this regard, the implementation of UASs in the construction industry as a project management tool is being explored by licensed remote pilot researchers in CACIM, as well as in collaboration with industry professionals. One such research study explored the feasibility of utilizing UASs in active construction environments (see Fig. 13) [15]. Primarily this was completed through comparing the potential use of UASs to the utilization of standard aerial imaging defined through a survey of construction industry professionals. Possible ways a project manager may utilize such technology for everyday operations, possible applications for the captured image and video data, as well as the financial impacts of employing a UAS were compared and contrasted to the existing industry practices. Further concerns regarding the use of UASs in an active construction environment as well as FAA guidelines were also analyzed. This research offered valuable feasibility of use insight for researchers and industry professionals as they consider how to best utilize UASs in the construction industry.

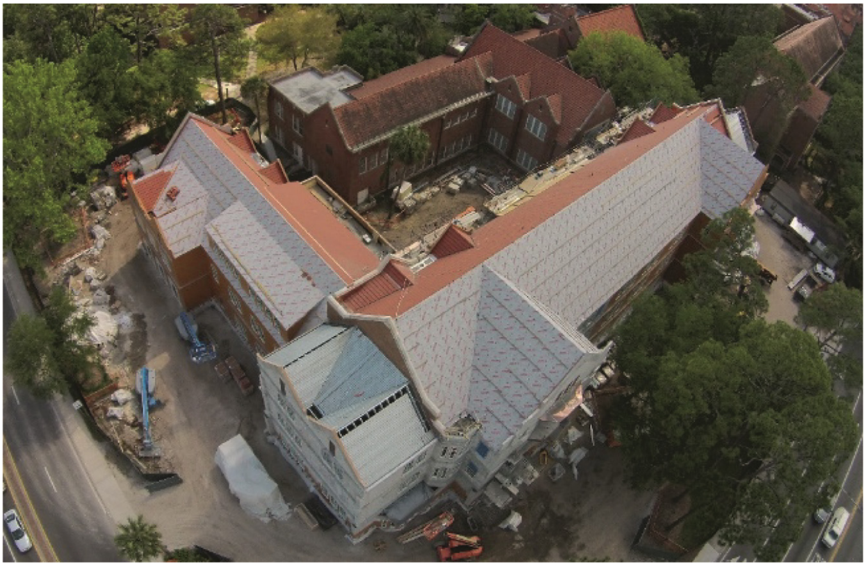


Fig. 13. Construction progress image captured via an UAS.

The emergence of UAS technology has spurred research to discover ways it could be implemented and used to make aspects of construction safer or more efficient. While many are looking into UAS technology there has been very little shared data on the state of aerial imaging practices in the construction industry. This research was conducted to establish a baseline of understanding of the current state of aerial imaging in the construction industry and provide a feasibility assessment for the use of UASs. Survey data indicated that the majority of the aerial imaging completed in the construction industry is captured on a monthly basis through the use of a third party manned-aircraft flyover. It was determined that feasibility should be established comparing the use of UASs to the most common existing practice, monthly aerial image collection using a third party manned-aircraft flyover. The final results indicate that the use of an UAS is less costly and more financially feasible than the current aerial imaging methods in use.

This research by CACIM was an industry leading step forward towards a thorough understanding of UASs and their place in the construction industry. This research began in the very early years of UASs development and researchers in CACIM have striven to remain on the forefront of UASs practices. Through a focused effort to remain in front of new regulations and lead forward through compliance and evaluation, CACIM has continued to be among the top research centers in the country conducting UAS research for construction operations. As the construction industry continues to evolve, new technologies will be employed to keep up with the increasing demand for efficiency, quality and safety. While the widespread use of UASs is still uncertain due to legal constraints, it is clear that UASs provide a feasible option for project teams to collect aerial images of their jobsites and have a variety of additional applications under development. While there is still work to be done, the research conducted by CACIM researchers demonstrates the feasibility of UAS implementation for construction management applications when compared to current industry practices and provides a benchmark for future studies.

2.6 Educational Video Games

The increasing demand for a skilled construction workforce that is capable of managing complex projects has pushed construction academic programs to their limits. In addition to accommodating the emerging new technologies and materials used in new projects, these academic programs are continuously trying to close the gap between the academic field and the professional one [16]. Almost all students who graduate with a construction management degree have at least the minimal required technical skills to perform their jobs. However, new graduates are faced with a steep learning curve when it comes to applying the knowledge transferred to them via the curriculum in the field.

Accordingly, academic institutions have sought to provide as much field experience as possible to their students via internship opportunities in order to produce better graduates that can handle the required job in the shortest period possible. This is one of the principle reasons why construction education should look at new techniques that would utilize existing and future technologies to improve the experiential learning of their graduates. Gaming environments have the potential to help these academic institutions accomplish their goals.

The existing educational system follows an outdated method of breaking a topic down to sections and introducing these sections one at a time, in serial fashion to students. The students are then expected to understand and memorize that information in order to pass a test. Only after the test is administered will the students be introduced to the next part of the topic. Usually, these parts build on each other to form the system of topics. However, students often lose the connections between these bits and pieces of the topic covered and therefore lose the big picture that forms the system [17].

Creating the proposed game consisted of three main steps: design, production and testing. The focus of the designed game, named Construction Virtual Experience (CONVEX), is to introduce construction management concepts to players in a fun and engaging manner. When individual enjoyment and engagement are high, players do not even realize that they are participating in an educational experience. Instead, players focus on the tasks at hand and try to surpass the high point scores of previous players.

The lack of published video games that teach construction management concepts in a true gaming environment lead to the decision to design CONstruction Virtual EXperience (CONVEX). The focus of CONVEX is to introduce construction management concepts to players in a fun and engaging manner. When the enjoyment and engagement are high, players do not even realize that they are being subjected to an educational experience. Instead, players focus on the tasks at hand and try to surpass the high point scores of previous players.

The best genre of video games to represent the construction industry is simulation-strategy. Construction projects are a series of activities that require the use of finite, budgeted resources to be completed without neglecting the important factor of time. Simulation-strategy games follow a very similar approach where players try to manage certain resources and compete against other players on the basis of completion duration. The faster players build their camp and expand, the higher their chances to win the game and beat their competitors. The same concept is used to design CONVEX except that there are no other real-time opponents to play against. Instead, CONVEX players are in constant battle with time since each activity is designed to start and finish in certain time periods. Figure 14 shows the overall workflow of the game.

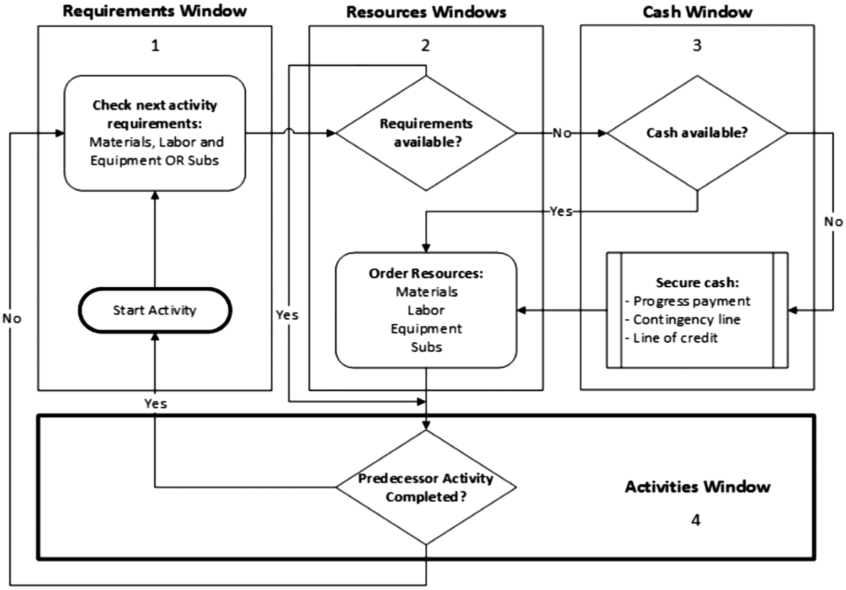
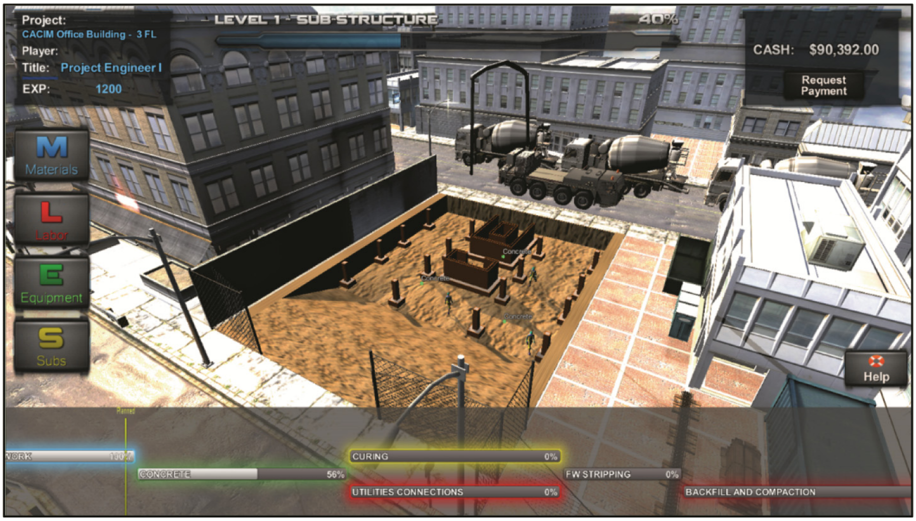

Requirements Window
1
Check next activity requirements: Materials, Labor and Equipment OR Subs
Start Activity
Resources Windows
2
Requirements available?
Order Resources: Materials Labor Equipment Subs
Cash Window
3
Cash available?
Secure cash: - Progress payment - Contingency line - Line of credit
No
Yes
No
No
Yes
Yes
Predecessor Activity Completed?
Activities Window
4

Fig. 14. CONVEX general workflow

The game is designed to keep players constantly engaged by continuing to check the resources for upcoming activities. However, there is a thin line between engagement and frustration. The game has to keep that balance to avoid losing players due to frustration. In order to minimize frustration, the game needs to start slow and give players a chance to get familiar with the different elements and requirements to complete it. After that, the pace needs to pick up to keep players engaged. Once players figure out the cycle of the game, it becomes easier and more entertaining.

Although it is very important for the game to be educational, that should not overtake the entertainment aspect. CONVEX is designed in a way that players are not constantly reminded about the educational factor of the game. Instead, the game relies on the engagement and entertaining factors to drive the educational one.

The game was developed using the Unity3D [18] gaming engine using realistic three-dimensional objects that represent actual building components and construction equipment. Most of the building components in the game were imported from Autodesk Revit through Autodesk 3ds-Max. All other components in the game were acquired from the Unity3D Asset Store. This significantly cut down development time since very few 3D modeling and animation needed to be done. Figure 15 shows a screenshot of CONVEX, which was developed to be solely controlled though the graphical user interface (GUI). That means players cannot interact with the 3D objects on the screen but rather control them by clicking the appropriate GUI elements. The design and development of the GUI system has gone through several versions to make it easier for players to locate the needed information and act upon it. Shanbari and Issa [19, 20] discussed the different factors that were taken into account when designing and developing the GUI system for CONVEX.

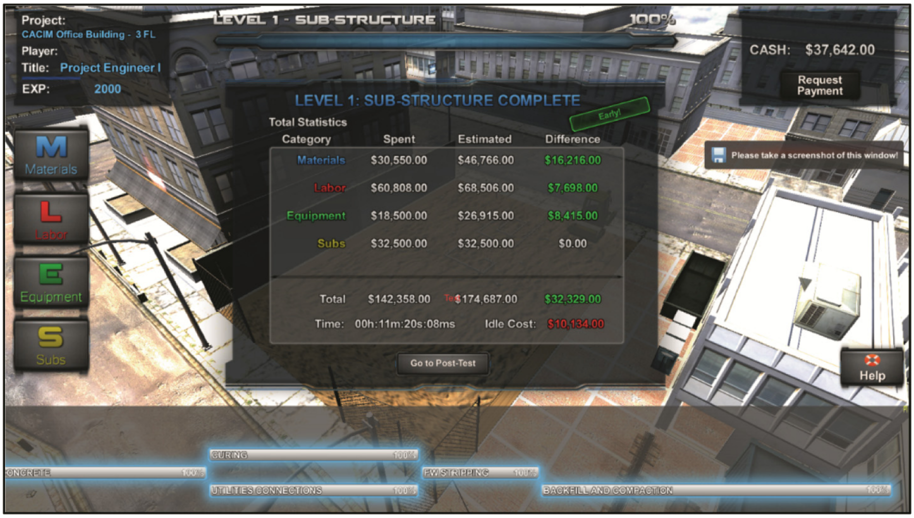


Project:
CACIM Office Building - 3 FL
Player:
Title: Project Engineer I
EXP: 1200

CASH: $90,392.00

Fig. 15. Screenshot of CONVEX gameplay (Color figure online)

Part of the development goals was to reduce the level of frustration in the game. That was accomplished in three ways. The first way was to introduce the players to the different GUI elements on the screen via a dedicated tutorial window. This allowed players to identify the different elements on the screen and know where to find certain information. The second way was to reduce player guessing by highlighting the elements that needed to be clicked or interacted with to move forward in the game. This can be seen in Fig. 15 where the activity bars are highlighted in different colors. These colors represent the status of each activity bar: red means resources are missing; yellow means resources are available but awaiting predecessor activity to finish; green highlights in-progress activities; and blue indicates a finished activity. The third and final way of reducing frustration is the help button. Using a complex algorithm, the help button, once clicked, determines the current status of the project and alerts the player of what step to take next. Figure 16 shows the completion window of the first level.



Project: CACIM Office Building - 3 FL

Player:

Title: Project Engineer I

EXP: 2000

CASH: $37,642.00

LEVEL 1: SUB-STRUCTURE COMPLETE

Total Statistics

Category	Spent	Estimated	Difference
Materials	$30,550.00	$46,766.00	$16,216.00
Labor	$60,808.00	$68,506.00	$7,698.00
Equipment	$18,500.00	$26,915.00	$8,415.00
Subs	$32,500.00	$32,500.00	$0.00
Total	$142,358.00	$174,687.00	$32,329.00

Time: 00h:11m:20s:08ms Idle Cost: $10,134.00

Please take a screenshot of this window!

Fig. 16. Screenshot of CONVEX level completion window (Color figure online)

CONVEX introduced students of various backgrounds to construction projects as a whole and a system where every component affects the others. The results of the study indicated that students from different majors and backgrounds had knowledge about construction management concepts comparable to those majoring in construction management after playing the game. These results indicated that high-level construction management concepts can indeed be taught using video games.

The main attribute that made a significant impact on the educational benefit of CONVEX was completing the first level of the game, which was developed for the test case. Players who managed to complete the game scored an average total score on the assessment test of 27% higher than those who failed to complete it. The group that failed to complete this first level indicated in the feedback that they had gotten confused at some point during the game. This confusion might have led to frustration and therefore quitting or failing (running out of time or cash before completion) this level. Such confusion and frustration can get the players unfocused and therefore result in their loss of some of the educational aspect of the game.

CONVEX introduced students of various backgrounds to construction projects as a whole and a system where every component affects the others. The results of the study indicated that students from different majors and backgrounds had knowledge about construction management concepts comparable to those majoring in construction management after playing the game. These results indicated that high-level construction management concepts can indeed be taught using video games.

The main attribute that made a significant impact on the educational benefit of CONVEX was completing the first level of the game, which was developed for the test case. Players who managed to complete the game scored an average total score on the assessment test of 27% higher than those who failed to complete it. The group that failed to complete this first level indicated in the feedback that they had gotten confused at

some point during the game. This confusion might have led to frustration and therefore quitting or failing (running out of time or cash before completion) this level. Such confusion and frustration can get the players unfocused and therefore result in their loss of some of the educational aspect of the game.

3 Conclusion

In addition to conducting advanced research, CACIM strives to continuously improve the education experience of all students in the Rinker School of Construction Management in the area of construction technology. In this regard, the BIM/VDC curriculum is continuously under development and improvement of, with a focus on process and conceptual development enabled through technical skills. Such experiences prepare them for their eventual careers and are an important part of ensuring that the industry as a whole continues to move forward. Pedagogical techniques implemented by CACIM delivered courses include an emphasis on experiences founded in industry practice, real-time collaborative exercises, technology exposure, and most importantly the belief that a student needs to understand the workflow implication and reasoning for any given toolset in order to facilitate its use. In addition to the curriculum based course work, CACIM also trains a competition team of undergraduate students who very successfully compete in a national VDC competition each year. This puts the industry's most advanced toolsets in the hands of future construction managers and provides an invaluable experience as they look to begin their careers.

At its core, CACIM is a research center focused on the advancement of construction technologies with the goal of facilitating meaningful progression toward a more efficient and effective industry as a whole. Advanced construction technologies provide never before seen data streams to project teams and enable highly effective data-driven decision making processes throughout the life of a project. A passion for the AEC industry, construction technology, student learning, and the Rinker School of Construction Management is a pre-requisite to be a part of the team and truly permeates each member's work. Cutting edge technology research, industry integration, and strategic development are a part of the everyday work conducted by CACIM and will be a major part of the AEC industries future. Lessons from the past are an important part of moving forward, and at CACIM the exploration of the path forward drives all of the research, education, and collaboration which makes it an industry leading academic center.

In closing, the purpose of this presentation is to describe the various approaches taken to collaborative knowledge sharing, technology innovation, knowledge transfer and integration into company workflows. Furthermore, the successful integration of these advances into construction curricula will help prepare students for an ever changing world of construction information technology.

References

1. Shanbari, H., Blinn, N., Issa, R.R.A.: Introducing laser scanning technology in a graduate BIM class. In: Issa, R.R.A. (ed.) Proceedings 9th BIM Academic Symposium and Job Task Analysis. pp. 183–190. Washington, DC (2015)
2. Blinn, N., Issa, R.R.A.: Academia meets AECO industry reality: a collaborative BIM coordination exercise in AECO education. In: Issa, R.R.A. (ed) Proceedings 2018 BIM Academic Symposium Orlando, FL (2018)
3. Blinn, N., Issa, R.R.A.: Utilization of drawing management software to enhance BIM educational experiences. In: Issa, R.R.A. (ed.) Proceedings 2017 BIM Academic Symposium. Boston, MA (2017)
4. Blinn, N., Issa, R.R.A.: Enhancing BIM educational experiences with integrated keystroke capture software. In: Issa, R.R.A. (ed.) Proceedings 2016 BIM Academic Symposium, pp. 102–09. Orlando, FL (2016)
5. Rankohi, S., Waugh, L.: Review and analysis of augmented reality literature for construction industry. Vis. Eng. **1**(1), 1–18 (2013)
6. Chi, H.-L., Kang, S.-C., Wang, X.: Research trends and opportunities of augmented reality applications in architecture, engineering, and construction. Autom. Constr. **33**, 116–122 (2013)
7. Shanbari, H., Blinn, N., Issa, R.R.A.: Using augmented reality video in enhancing masonry and roof component comprehension for construction management students. Eng. Constr. Architect. Manag. **23**(6), 765–781 (2016)
8. Blinn, N., Robey, M., Shanbari, H., Issa, R.R.A.: Using augmented reality to enhance construction management educational experiences. In: Proceedings 32nd CIB W078 Workshop. Eindhoven, The Netherlands (2015)
9. Bademosi, F., Blinn, N., Issa, R.R.A.: Use of augmented reality technology to enhance comprehension of steel structure construction. In: Proceedings Lean & Computing in Construction Congress (LC3), vol. 1, (CIB W78). Heraklion, Greece (2017)
10. Shanbari, H., Blinn, N., Issa, R.R.A., Robey, M., Zhu, Y.: Using augmented reality video in enhancing roof components comprehension for construction management students. In: Proceedings ICCBE International Conference on Computing in Civil and Building Engineering, pp. 1915–1922. Osaka, Japan (2016)
11. Blinn, N., Tayeh, R., Issa, R.R.A.: Visualizing and coordinating construction projects using holography. In: Proceedings 2018 Construction Congress. New Orleans, LA (2018)
12. Shanbari, H.A., Blinn, N.M., Issa, R.R.A.: Laser scanning technology and BIM in construction management education. In: ITcon 21(0) Special Issue 9th AiC BIM Academic Symposium & Job Task Analysis Review Conference, pp. 204–217 (2016)
13. Blinn, N., Issa, R.R.A.: Comparison of traditional phase shift laser scanning and infrared scanning techniques. In: Proceedings ASCE International Workshop on Computing in Civil Engineering. ASCE, Seattle, WA (2017)
14. Blinn, N., Issa, R.R.A.: Parametric model for planning laser scanning jobs in the AECO industry. In: Proceedings ICCBE International Conference on Computing in Civil and Building Engineering, pp. 1468–1475. Osaka, Japan (2016)
15. Blinn, N., Issa, R.R.A.: Feasibility assessment of unmanned aircraft systems for construction management applications. In: Proceedings 2016 ASCE Construction Congress, pp. 2593–2603. San Juan, PR (2016)
16. Hasan, H.S.M., Ahamad, H., Mohamed, M.R.: Skills and competency in construction project success: learning environment and industry application- the gap. Procedia Eng. **20**, 291–297 (2011)

17. Yves, B.: Contemporary Theories and Practice in Education. Atwood Publishing, Madison (2003)
18. Unity 3D [Computer Software]: Unity Technologies. San Francisco, CA (2016)
19. Shanbari, H., Issa, R.R.A.: Graphical user interface in interactive education: an implementation in the construction industry. In: Proceedings CONVR. Banff, Canada (2015)
20. Shanbari, H., Issa, R.R.A.: Use of video games to enhance construction management education. Int. J. Constr. Manag. (2018). https://doi.org/10.1080/15623599.2017.1423166

Derivation of Minimum Required Model for Augmented Reality Based Stepwise Construction Assembly Control

Nicolas Jeanclos[1(✉)], Mohammad-Mahdi Sharif[1], Shang Kun Li[2], Caroline Kwiatek[1], and Carl Haas[1]

[1] Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada
njeanclo@uwaterloo.ca

[2] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada

Abstract. With the advancements in 3D-imaging technology, new quality control methods can be developed and applied in the construction industry. Strictly stipulated tolerances put contractors under pressure to assure a high level of quality and accuracy. Thus, tolerance control has become imperative in the construction industry. Industrial assemblies are usually fabricated through stepwise processes which are being automated at a rapid pace. New components are added to partially-built assemblies composed of one or more components. Consequently, for each step, geometrical control can be applied to prevent any ongoing deviation from being propagated throughout the completion of the assembly. To do so, an intermediate 3D-model is derived from the initial design CAD file. Utilizing the Minimum Required Model (MRM) as opposed to the whole model enables easier quality control when using a 3D-imaging device and makes the control of the orientation and alignment faster, more accurate, easier and safer. This paper proposes a novel method to derive the MRM for stepwise quality control of a construction assembly using 3D-imaging technologies. Developed for use in the piping industry, the method uses the complete 3D-model and the Piping Component File (PCF), which includes the overall assembly in a generic text file format, to reduce the boundaries of the model based on the step currently being assembled. This process saves time by reducing the volume of the required scan to be processed and is more accurate than using manual tools for measuring the alignment. The algorithm for the derivation of the MRM is evaluated on 95 different scenarios and the results are compared in terms of reduced level of complexity and reduced principal length.

Keywords: 3D model · Optimization · Quality control · Augmented reality
Fabrication process · Mobile scanning

© Springer International Publishing AG, part of Springer Nature 2018

I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 336–358, 2018.
https://doi.org/10.1007/978-3-319-91635-4_18

1 Introduction

The emergence of modularization and prefabrication in the construction industry has enhanced the industry's performance. Modular construction reduces duration of the construction phase, saves on construction costs and improves end-product quality [1]. Given the controlled nature of the factory environment, quality control can be performed more accurately, and contractors are held to a higher standard to achieve strict tolerances. As a result, construction companies have looked to other manufacturing industries – such as the automotive industry – and started to use their measurement processes by adapting them [2].

Automation of construction processes in general and modular assembly specifically, would enable contractors to achieve a higher degree of accuracy with respect to geometrical deviations. For instance, automated machines are developed and utilized for welding and cutting [3].

Due to the nature of construction industry, most of the work is still manual and labor intensive (unlike the automotive industry). Furthermore, in the U.S, the construction industry has the highest fatality rate across all industries with approximately a thousand deaths in 2015 [4]. The high level of skill required coupled with the high accidents rate has led to a major challenge in supplying the required workforce in construction projects [5]. Focusing on the piping industry, workers need to develop a variety of skills, including, interpretation of spatially complicated assemblies from two-dimensional isometric drawings, correctly positioning components and proper welding.

In the piping industry, design files are transmitted in the form of a 2D drawing. These drawings are then printed and distributed to workers on the fabrication floor. Workers are trained to use these 2D drawings to build the 3D model of the design in their minds and then lay out the pipe spools based on their interpretation. Because of the complexity of this task, workers usually develop their own method to better understand (visualize) the drawing. One such method is to use a welding rod to quickly build a 3D model of the assembly. As such, utilizations of 3D models (which could be digital or 3D-printed) along with the traditional 2D drawing has been proven to be effective when assembling components together [6]. However, utilization of 3D models in the prefabrication stages has been very limited due to the substantial changes in the internal workflows that it requires.

One of the major bottlenecks in modular fabrication is the discontinued inspection process. Be it in the form of utilization of high end laser scanners, experienced Quality Control (QC) personnel or third party QC firms, they can only be applied at the end of the fabrication process. Moreover, detection of a geometric deviation at the end of the fabrication process can cause major delays and costs to the project [7]. Recently developed range cameras can potentially be a solution to this problem since they can offer accurate real time information on the as-built status that can be then compared with the 3D model [8]. In addition to their ability to collect real time information, these scanners require a minimum training experience, making them a great solution to be adapted on the fabrication floor. However, range cameras have a limited range of acquisition which is the primary reason of their limited application in the industry [9]. To address this challenge, this paper offers a method to reduce the percent of the 3D model required to be scanned by introducing the novel concept of Minimum Required Model (MRM). By deriving the MRM, the challenge created by the short range of these sensors will be circumvented, and a continuous feedback loop for the workers can be developed and utilized to avoid any deviation from the design in the fabrication stage.

The following paper is organized in three section. To begin, the background of this research is presented and summarized, including the existing literature in quality control in prefabrication construction facilities, utilization of 3D models in fabrication processes and some 3D vision technologies for automated inspection in fabrication. In the next section, the methodology is developed, and the concept of MRM is extensively explained. To validate the effectiveness of the MRM, the results are then provided. Finally, conclusions and future work are detailed in the last section.

2 Background

This section is broken down to four sections: (1) an overview of the current quality control processes in pipe fabrication and their impact on obtaining the 3D model of the design, (2) utilization of 3D vision technologies for visual inspection, (3) novel imaging tools and techniques to improve QC in construction industry, and (4) knowledge gap and problem statement of the proposed paper.

2.1 Quality Control Processes in Pipe Fabrication

The current market of the U.S pipe fabrication industry in 2017 is evaluated at $45 billion [10]. Previous researchers [11] have shown up to 10% of construction costs are attributable to rework due to late defect detection, among which 50% is caused by human errors [12]. This would yield $2 billion as the amount that could potentially be saved in pipe fabrication industry by early defects detection.

The predominant processes for monitoring the geometrical correctness of fabrication assemblies in the piping industry involves manual inspection by certified QC personnel using direct contact measurement devices, such as, metal measuring tapes, calipers, custom gauges, squares, and straight edges. These manual methods of inspection can generate measurement uncertainties, which have to be taken into account when verifying assemblies' compliance. Uncertainties and errors in measurement can come from many sources, such as the measuring device, the component being measured, the skill of the craft worker or inspector performing the measurement, the measurement process, and the environment.

In addition to measurement uncertainty, dimensional control problems also originate from poor existing document creation practices. For example, chain dimensioning establishes ambiguity and the potential for accumulated measurement error or the drawings include units that may not be legible to the fabricator which may cause delays.

Another problem is the fragmented process from the design stage to the fabrication. Although not universal, some of the major steps include: (1) 3D design and modeling, (2) transmitting isometric drawings, (3) production drawing creation (referred to as "cutsheets"), and (4) work package distribution and fabrication. Depending on the contract, a different company may do each of these steps. As such, a major challenge is to address the interoperability issues, such as, different software packages used in the modeling process [13]. Companies have developed different strategies as how to address the interoperability issue, including using in-house workforce to re-model the design using 2D drawings, which can be very costly and inefficient [14] or, in many projects, not taking advantage of the 3D models. However, 3D modelling and especially Building Information Modelling (BIM) have been proven to be of a major importance in construction projects. Communication, coordination, visualization and clash detection are among the major advantages BIM models [15, 16]. As a result, in order to remedy the interoperability challenge and keep the benefits of using 3D-models, software independent CAD file formats have emerged. For instance, a commonly used format is Stereolithographic format (*.Stl) [17] which is compatible with a variety of software packages and only contains the geometrical information of an assembly [18].

Furthermore, a file data transfer method is used in the piping industry to pass along the most important geometrical information and management information related to the pipe spool assembly. Piping Component File (PCF) is a type of file which contains this important information [19]. PCFs are formatted as a text file and hence are not software dependent. End-points of the components, diameters, length, type of weld, and specific description are some instances of the PCF, which is segmented in paragraphs specifying each component or weld.

2.2 Automated Visual Inspection in Construction

Automated inspection is desirable because manual inspection by humans is time-consuming, and can be excessively subjective, unreliable, and not interesting for humans to perform. Also, many industrial assemblies are not easily accessible for manual inspection. Emergence of CAD models in the mid 1990's [20] and advanced 3D imaging technologies allowed for accurate dimensional evaluations as automated parts of the fabrication process. Bosché et al. [21] presented a methodology for using 3D laser scanners and comparing the as-built data with the BIM model for construction projects.

Building on this methodology, Nahangi et al. [22] presented an automated approach for monitoring and assessing fabricated pipe spools and structural systems using automated scan-to-BIM registration. The method reliably detects the presence of dimensional non-compliance and has consistently quantified deviations with less than 10% error in experimental studies. The method requires two 3D imaging input files: (1) a point cloud of the as-built assembly generated using a 3D reconstruction technique such as Light detection and ranging (LiDAR), and (2) the tolerance specifications as represented by a 3D CAD design file. LiDAR, often referred to as laser scanning in the construction industry, is an increasingly important technology. However, the acquired data is too dense and a number of methods have been developed to automatically find objects of interest(s) [23]. Similar methods include the combination of BIM and LiDAR to build a real-time construction quality control system [24]. It is the state-of-the-art that, 3D laser scanning technologies are commonly used in construction to inspect the quality of a project [25, 26].

Researchers have created automated methods for monitoring and performing automated 3D image-to-BIM comparison of mechanical, electrical, and plumbing (MEP) systems [21, 28], and general building and structural systems [26, 29–31].

Using 3D imaging for dimensional compliance assessment of construction components has the potential to mitigate costly repair and rework while tracking progress [32].

Laser scanning technologies have thus been under extensive investigation and have been successfully applied in the industry [31]. Yet, most of the time, laser scanners are merely used at the end of the fabrication process to control and validate the assembly, because 3D imaging technologies require engineering work and are usually labor-intensive [12]. Consulting companies have to come in the fabrication facility to scan the assembly and then have to process the data before providing a final report. The process is expensive and takes several days [33]. Since the data captured by laser scanning typically includes a large number of point clouds, the derivation of as-built models in a quick and cost-effective manner is a major challenge [34–36].

2.3 New 3D Vision Technologies in Fabrication

With the advancements of technology, new methods have emerged such as augmented reality (AR) systems. Combined with BIM, AR can provide a full 3D interactive solid model of the design, giving the workers a visual understanding of the design. Wang et al. [37] describe some of the AR functionalities such as superimposing a 3D colored model onto the standard 2D drawing plan. As information can be made available in real-time, AR can be used to expedite tasks efficiently and effectively [38]. Head mounted displays (HMDs) can potentially be fully integrated in QC processes [39, 40]. In [41], a welding helmet is presented which provides augmented visual information such as drawings and quality assistance. Schwerdtfeger et al. [42] employ laser projectors for stepwise fabrication. Laser beams of the component's boundaries are projected on the surface where it is to be assembled. Molleda et al. [43] developed a 3D imaging system for dimensional quality inspection that utilizes range images. Another category of acquisition devices for collecting as-built information is Unmanned Aerial Vehicles (UAVs). Siebert and Teizer used an UAV system to survey earthwork of large infrastructure projects [44]. Other uses include and are not limited to project tracking, quality inspection, and safety inspection [45].

2.4 Knowledge Gap and Problem Statement

Recently, depth cameras have emerged as potential 3D imaging acquisition tools, but due to their short scanning range, no realistic utilization has been developed for construction assemblies [46]. Despite their range that for most of them does not exceed four meters, range cameras are inexpensive and potentially accurate. This means that workers can use them with little fear of breaking them and that a precise quality control and geometric compliance can be performed [47]. Combining the two characteristics opens the path to stepwise quality control of construction assemblies. However, the range constraint has to be addressed in order to be able to use range cameras for construction assemblies.

Construction assemblies are often larger than the intrinsic depth camera's range of a few meters [48]. To face this major challenge, the solution proposed in the following paper is to shrink down the dimensions of the required scan by redefining the boundaries of the model. The 3D model is derived to contain the minimum required geometrical information to enable the geometric control of the assembly. A method is developed to automatically derive the Minimum Required Model (MRM) and allows range cameras to be effectively used for construction assembly control. To achieve the derivation of the MRM, an assembly guidance process is designed for workers and facilitates the stepwise tracking of the assembly.

The proposed method combines STL and PCF file formats to prevent any possible interoperability challenges, to improve the control guidance and the quality control (QC) of the assembly in a stepwise manner. As new components are added to partially-built assemblies composed of one or more components, the geometric control can be applied at each step to prevent any ongoing deviation from being propagated throughout the completion of the assembly.

3 Methodology

This paper presents an effective method to derive the Minimum Required Model (MRM). Its utility consists in reducing the number of components to be scanned for a 3D vision quality control. The determination of these components is achieved to prevent any geometrical ambiguity in the MRM; that is, by ensuring the final derived assembly geometry, the whole assembly compliance is guaranteed.

The following section describes thoroughly the entire process of MRM derivation. As shown in Fig. 1, the process is divided into three different sections. In Sect. 3.1, the basic concepts of the method are explained; that is, explaining the use of Piping Component Files as well as a description of the appropriate categorization of components into Reference (REF) and Addition (ADD). In Sect. 3.2, the concept of Solid of Revolution (SOR) is presented and the determination of REF/ADD in SOR is explained. Finally, Sect. 3.3 explains how the MRM is derived using REF and ADD components.

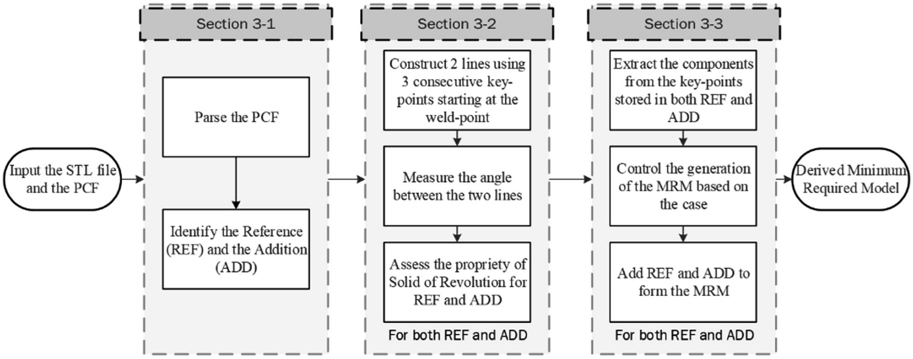

Section 3-1
Section 3-2
Section 3-3
Input the STL file and the PCF
Parse the PCF
Identify the Reference (REF) and the Addition (ADD)
Construct 2 lines using 3 consecutive key-points starting at the weld-point
Measure the angle between the two lines
Assess the propriety of Solid of Revolution for REF and ADD
For both REF and ADD
Extract the components from the key-points stored in both REF and ADD
Control the generation of the MRM based on the case
Add REF and ADD to form the MRM
For both REF and ADD
Derived Minimum Required Model

Fig. 1. Minimum required model workflow

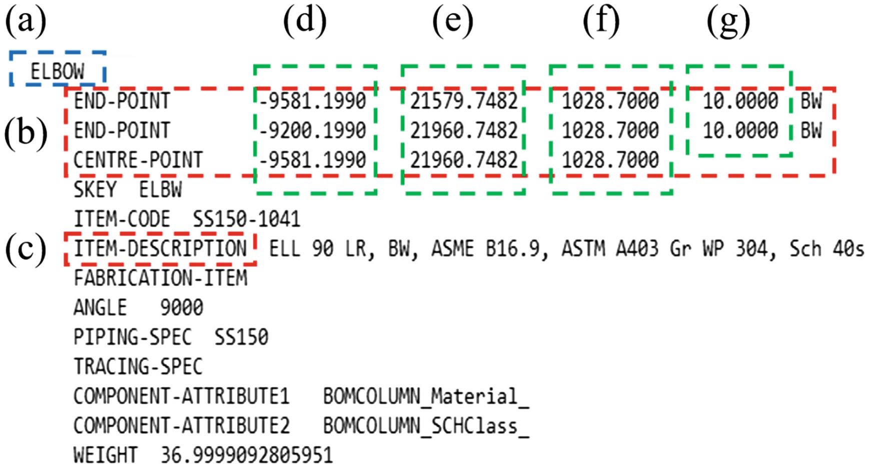
3.1 Parsing of the Piping Component File and Selection of the Reference and the Addition

A major aspect of the MRM algorithm is its use of Piping Component Files (PCFs) and Stereo-Lithography files (STLs). Combining the 3D model, in the format of a non-software dependent file such as a STL along with a PCF makes the method completely software independent. It solves the software compatibility challenge encountered by construction companies [49]. The first part of this section details how a PCF is parsed to extract the appropriate information. The identification of Reference and Addition is explained afterwards.

Parsing of the Piping Component File

As discussed in the background, a common practice in the piping industry is to transmit the PCF of the assembly along with the 2D drawings and/or the 3D model.

Transmitting the PCF can be seen as a safe means to send the drafting information, because it contains all the components' information including their type, coordinates, size, shape, and weld locations. The file is segmented into paragraphs where each one represents a different component. Figure 2 shows an example of the information which can be found in a PCF. Each paragraph has the following attributes: (a) the name of the component, (b) the key-points, (c) a precise description, and (d) the key-points coordinate and diameter. The combination of the key-points allows a complete description of the component. For example, a pipe has two key-points while an elbow needs three key-points to be modelled. A representation of the most common components with their associated key-points – a key-point can either be an end-point, a center-point or a branch-point – is provided in Fig. 3.



```
ELBOW
    END-POINT           -9581.1990  21579.7482  1028.7000  10.0000  BW
    END-POINT           -9200.1990  21960.7482  1028.7000  10.0000  BW
    CENTRE-POINT        -9581.1990  21960.7482  1028.7000
    SKEY  ELBW
    ITEM-CODE   SS150-1041
    ITEM-DESCRIPTION  ELL 90 LR, BW, ASME B16.9, ASTM A403 Gr WP 304, Sch 40s
    FABRICATION-ITEM
    ANGLE   9000
    PIPING-SPEC  SS150
    TRACING-SPEC
    COMPONENT-ATTRIBUTE1   BOMCOLUMN_Material_
    COMPONENT-ATTRIBUTE2   BOMCOLUMN_SCHClass_
    WEIGHT  36.9999092805951
```

Fig. 2. An example Piping Component File: (a) the described component is an elbow, (b) three key-points: two end-points and one center-point, (c) item description where, for instance the piping standard is specified, (d) x-coordinates of the key points, (e) y-coordinates of the key-points, (f) z-coordinate of the key points, (g) diameter about the key-points.

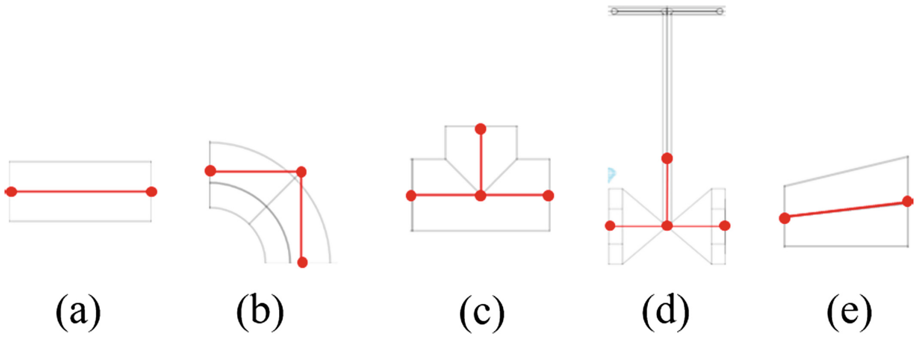


Fig. 3. Demonstration of piping components using key points and the hypothetical line to represent each component: (a) a pipe or any other cylindrical component with 2 end-points, (b) an elbow with two end-points and one center-point, (c) a tee with two end-points, one center-point and one branch-point, (d) a valve with two end-points, one center-point and one branch-point, and (e) a reducer-eccentric with two end-points.

Using the key points, an envelope is calculated and encompasses the corresponding component in the .STL file. Thus, the envelope provides the possibility to segment a specific component out of the assembly. Employing PCFs would allow users to only select and segment the components they are working with as opposed to the complete assembly.

Identification of Reference and Addition

The derivation of the Minimum Required Model is to be integrated with a stepwise process of fabrication, meaning that workers would control the compliance of the fabricated assembly in each step. One of the important steps for pipe fitters is to choose the sequence of the assembly. The selection of the sequence depends on a number of variables including maximizing number of roll welds. Regardless of the sequence, there would always be a component, or group of components, taken as the Reference (REF) which would be leveled and the next component, or group of components, would be added, called the Addition (ADD). Figure 4 further illustrates the concepts of ADD and REF on an industrial assembly.

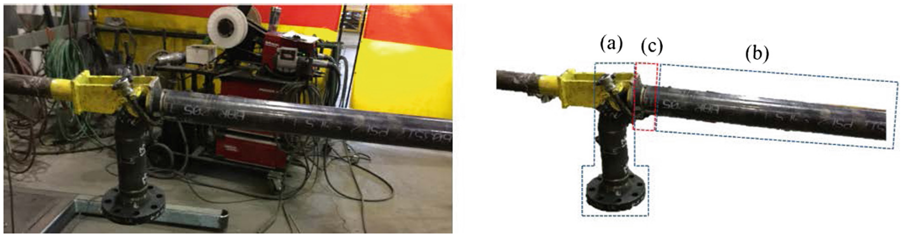


Fig. 4. Illustration of REF, ADD and connection point on a pipe spool assembly in a fabrication facility. On the left, a photo from a fabrication shop representing an assembly ready for welding. On the right, this assembly is analyzed for the purpose of the methodology. (a) Reference consists of a flange, a pipe, and an elbow. (b) Addition is a straight pipe. (c) The weld point shared between REF and ADD.

The sequence by which an assembly is built depends on the worker. For example, in Fig. 4, a worker could have decided to first weld the pipe in the box (a) to the elbow in box (b) and then weld the pipe in the box (a) and the elbow in the box (b) with the pipe in the box (b) and the flange in the box (b). In sequencing the components as REF and ADD, workers' objective is to maximize the number of roll welds. The proposed method is robust to any combination of REF and ADD since the algorithm takes advantage of the welding point between REF and ADD, and does not rely on the components available in each group. The weld-point accounts for the starting point of REF and ADD. It is worth noting that a weld-point is shared by two components and thus belongs to both REF and ADD. Also, it is the only shared point between REF and ADD (Fig. 4 box (c)).

For any given assembly, a graph is built using the key-points as nodes. The weld-point is used as the root of the graph. Once the weld point is found, its neighbors are then linked to each other.

Figure 5 illustrates the graph construction of an assembly that has pipes, elbows, tee and valves. The algorithm starts reading the key-points information from the weld-point and follows the flow of the assembly through the graph. As shown, the weld point is located between the blue part (REF) and the red part (ADD), between a pipe and an elbow. Considering the Reference graph (in blue), the weld moves on to the adjacent point, standing for the second end-point of the pipe. Then, it follows linearly the path with an elbow, a pipe, an elbow and another pipe. At that point, the first end-point of the tee is divided into two children: one is the branch-point and the other is the second key-point. The process will be terminated once all the components have been analyzed.

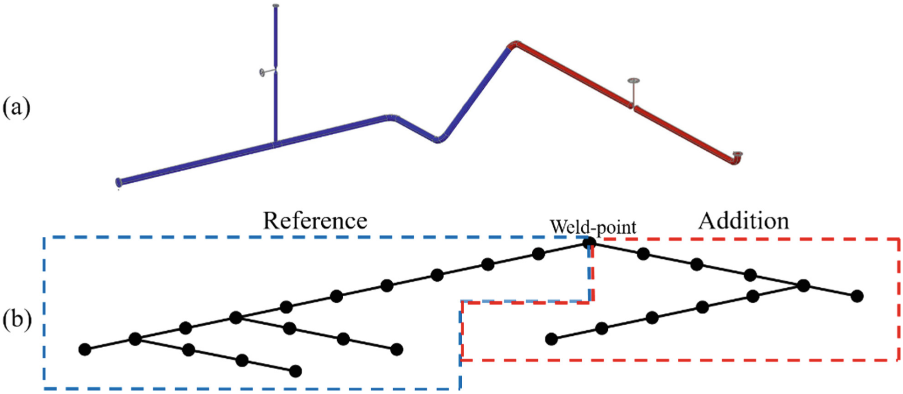


Fig. 5. Construction of the graph using the 3D model and the PCF file. (a) 3D model assembly. (b) Constructed graph. Each node corresponds to a key-point. (Color figure online)

Once the graph is built, REF and ADD have to be evaluated to determine whether or not they are a Solid of Revolution (SOR). This is the objective of the following section.

3.2 Determining the Property of Solid of Revolution for Both Reference and Addition

Solid of Revolution

In order to evaluate whether REF/ADD is a Solid of Revolution (SOR) or not ($\overline{\text{SOR}}$), a clear explanation has to be given around the concept of solid of revolution. By definition, a solid of revolution is an object that can be generated by rotating any arbitrary shape around a straight line [50]. With this definition, every cylindrical object (pipe), flanges, caps and couplings are a SOR. Elbows, tees, valves or reducer-eccentrics are not SOR ($\overline{\text{SOR}}$). Considering that piping components are assembled by aligning them together (sharing the same axis such as shown with $\overrightarrow{w_1}$ and $\overrightarrow{u_1}$ in Fig. 6), the SOR becomes a conservative property. This means that assembling two SOR piping components will result in a SOR assembly. Also, if a $\overline{\text{SOR}}$ piping component is added to a SOR component the assembly becomes $\overline{\text{SOR}}$. Equation 1 summarizes this concept:

$$\forall(S_1, S_2) \in \{\text{SOR}\}, \forall S_3 \in \{\overline{\text{SOR}}\}, \begin{cases} (S_1 + S_2) \in \{\text{SOR}\} \\ (S_1 + S_2 + S_3) \in \{\overline{\text{SOR}}\} \end{cases} \quad (1)$$

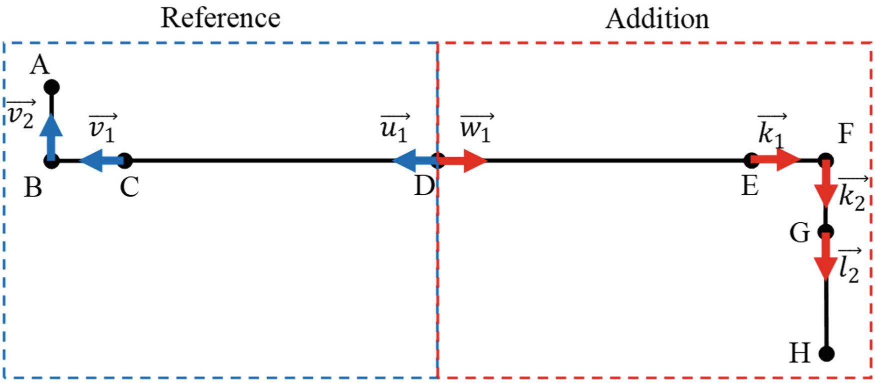


Fig. 6. 2D-representation of a Pipe-Elbow-Pipe assembly (Addition) welded onto an Elbow-Pipe assembly (Reference). Reference is composed of a pipe (CD) and an Elbow (CBA) and Addition is made of a pipe (DE), an Elbow (EFG) and a pipe (FH). The unit vectors represent the normalized vectors connecting two consecutive points.

where S_1 and S_2 denote to a piping component that is a solid of revolution, and S_3 denotes to a component that is not a solid of revolution. The next step is to assess the SOR property for REF and ADD in order to derive the proper Minimum Required Model.

Assessing the Solid of Revolution Property for Both Reference and Addition

Finding out the SOR property permits the creation of the key-points list for REF and ADD. The combined list will be used to extract the MRM. Algorithm 1 describes how this process is performed. The SOR property is defined as a Boolean variable, is initially set as true, and can be changed to false for specific conditions.

To assess whether a component is a solid of revolution or not (regardless if it is REF or ADD), the angle between two consecutive lines (using the key points) is evaluated. When a non-zero angle is detected, the two lines are not aligned and, hence, the group of components belongs to $\overline{\text{SOR}}$. If all angles are null or, in the case of a single component with two key-points, no angle can be calculated, REF or ADD is set as SOR. As written in Algorithm 1, the angle is evaluated between the current point P_i and its two neighbors P_{i-1} and P_{i+1}. Two normalized vectors are created binding P_i to its neighbors. The angle is then evaluated using the inverse cosine function between these vectors.

The detection of a non-zero angle reveals the presence of a $\overline{\text{SOR}}$ component in the inspected group of components. Following Algorithm 1, the explanation is either the current key-point possesses two children or it is not aligned with its previous and next neighbor. The first situation is detected when a tee or a valve is part of the assembly, whereas, the second situation happens when an elbow or a reducer-eccentric is in the assembly. For instance, if a 90-degree elbow is in the chain of components, when calculating the angle for its first key-point, an angle of 90° will be calculated. As a result, the whole group of components would be $\overline{\text{SOR}}$. For both Reference and Addition, the algorithm will be terminated as soon as a non zero angle is detected. In the situation where every key-point has been stored in the list (when REF and ADD are SOR), the algorithm will reduce the list to its initial configuration with only the weld-point.

Algorithm 1. Pseudo code for Solid of Revolution (SOR) determination and extraction of key-points for the Minimum Required Model

```
SOR=true
i=1
$P_0$ is the weld-point
KeyPointsList = {$P_0$}
while $P_i$ has at least 1 child
if $P_i$ has 2 children
  SOR=false
  break
else
  Calculate the angle between $\overrightarrow{P_{i-1}P_i}$ and $\overrightarrow{P_iP_{i+1}}$
  if (angle>0)
  SOR=false
  break
  end if
end if
add $P_i$ to KeyPointsList
i=i+1
end while
if all $P_i$ are in KeyPointsList
  KeyPointsList = {$P_0$}
end if
KeyPointsList
end
```

For both REF and ADD, the algorithm runs and their respective final key-points list, depending on the SOR property, is generated. Once the SOR property is determined, the final components of the MRM can be extracted. The next subsection focuses on the last step.

3.3 Derivation of the Minimum Required Model

As explained in the beginning of this section, the objective of MRM is to eliminate the geometrical ambiguities in the assembly in order to reduce the complexity of the assembly by reducing its scan volume and the number of components. Consequently, instead of analyzing and scanning the entire assembly to perform a geometrical control, the inspection can be done on the reduced assembly (MRM) only.

In the following section, the rules of component selection and the constitution of the Minimum Required Model are explained.

Analysis of the Structure for Both Reference and Addition

As explained in the previous section, a flow of vectors is constructed. Each vector is associated to a node and its direction is dictated by the component's direction in space. After being normalized, the unit vectors can be used to determine the value of the angle calculated (Algorithm 1). To facilitate the understanding, an example is provided in Fig. 6. The weld-point (D) is the initial position for both chains. From there, the unit vector representing the first pipe of REF is built ($\overrightarrow{u_1}$) between D and C. While going to the next point, the vectors are built in order to be able to calculate the angle. The process is then repeated for ADD.

It is crucial to understand that the geometric compliance of (DCBA) and (DEFGH) have been established in the hypothetical previous steps of fabrication. In other words, each sub-assembly is correct. Consequently, the step under investigation in Fig. 6 is the assembly of ADD (DEFGH) with REF (DCBA). When applying Algorithm 1, the key-points lists for REF and ADD are respectively composed of points D, C and B and points D, E and F.

Principle of Geometric Control

To assess the geometrical alignment, every axis of ADD has to be controlled with respect to the ones of REF. The objective is then to constrain the vectors from ADD to the ones from REF. The representation of vectors is provided in Fig. 6. To assess the geometric alignment, the vectors are compared together within each sub-assembly to only keep the ones needed. The process of isolating the unit vectors is done following the two rules of Rule 2 and Rule 3:

$$\text{If } \overrightarrow{a_1} = \overrightarrow{b_1} \text{ then only keep } \overrightarrow{b_1} \tag{2}$$

$$\text{If } \overrightarrow{a_1} \neq \overrightarrow{a_2} \text{ then keep } (\overrightarrow{a_1} \text{ and } \overrightarrow{a_2}), \tag{3}$$

where $\overrightarrow{a_1}$and $\overrightarrow{b_1}$ denote two unit vectors connecting two consecutive key-points. The analysis of vectors and their selection enables the constitution of the sub-set of points used to form the MRM and control the correctness of the assembly. Applying Rule 2 and Rule 3 to Fig. 6 would result in keeping $\overrightarrow{v_1}$, $\overrightarrow{v_2}$, $\overrightarrow{k_1}$, and $\overrightarrow{k_2}$ and removing all the remaining vectors.

To assess the alignment of ADD to REF, controlling the position of $\overrightarrow{k_1}$ with respect to $\overrightarrow{v_1}$ is required. By constraining them together, a revolute pair is created and removes 5 of the 6 degrees of freedom. In other words, after fixing $\overrightarrow{k_1}$, ADD has only one more degree of freedom left around $\overrightarrow{v_1}$. Therefore, an additional constraint must be set to remove the last degree of freedom. For this purpose, the position of $\overrightarrow{k_2}$ is fixed based on $\overrightarrow{v_2}$. ADD is thus completely constrained to REF. Finally, $\overrightarrow{l_2}$ does not have to be analyzed since ADD is fully constrained to REF.

However, in the situation where all the unit vectors are equal, for REF or ADD, then no constraint needs to be set up to remove the potential rotation around the first axis. Indeed, if REF has just one unit vector, there is no $\overrightarrow{v_2}$ to assess a potential $\overrightarrow{k_2}$; respectively, if ADD has just one unit vector there is no $\overrightarrow{k_2}$ to be constrained to a potential $\overrightarrow{v_2}$. The described situations stand for scenarios where either REF or ADD is a Solid of Revolution (SOR).

Constituting the Minimum Required Model

Once the key-points lists are finalized for REF and ADD (redundant points have been removed from the list), the constitution of the Minimum Required Model can be done. The outcome is to extract the components that are associated to these key-points. Because of the parsing of PCF, each key-point is directly associated to a component. Therefore, a component is retrieved when one of its end-points is detected in a key-point list. Once all key-points have been investigated, a minimum required Reference ($\mathrm{REF_S}$) and a minimum required Addition ($\mathrm{ADD_S}$) are created using the detected components. The Minimum Required Model is obtained by adding $\mathrm{REF_S}$ and $\mathrm{ADD_S}$. Equation 4 below provides the mathematical expression of the MRM once $\mathrm{REF_S}$ and $\mathrm{ADD_S}$ have been derived.

$$\mathrm{MRM} = \mathrm{REF_S} + \mathrm{ADD_S} \tag{4}$$

The derived 3D-model is only made up with the minimum required components to control the geometry and the alignment of the complete assembly.

Finally, four different situations can be encountered and impact the constitution of the Minimum Required Model. Table 1 summarizes the 4 situations which will be used to design the experimental 3D-models.

Table 1. All possible cases with respect to ADD/REF being or not being a SOR. The 1-value stands for SOR and the 0-value for $\overline{\text{SOR}}$.

Case	Reference	Addition	Outcome
1	1	1	MRM composed of the closest component to the weld of REF and ADD
2	0	1	
3	1	0	
4	0	0	MRM composed of components from the weld to the component associated to the non-zero angle for REF and ADD

4 Results

In order to validate the proposed methodology and to prove the effectiveness of the Minimum Required Model, a set of 95 scenarios is designed. Three performance measures are evaluated by comparing the initial models to their final MRMs: the spatial complexity (1) and the number of components (2) to be inspected before and after applying MRM, and the MRM's impact on the accuracy and the feasibility of collecting as-built information using consumer grade scanning devices (3).

The Minimum Required Model is analyzed on 95 scenarios, where each scenario represents a piping assembly. The assemblies are categorized into four possible cases presented in Table 1. All assemblies are designed using *AutoCAD Plant 3D 2018*. All the components belong to the same standard piping catalog. The scenarios are designed by the authors and have been approved by experts in the industry as realistic assembly models.

Case 1 as defined in Table 1 includes 14 scenarios. Case 2 and 3 contain 15 scenarios and case 4 is comprised of 51 scenarios. Five samples of the 3D-models for each case is provided in Fig. 7. An ID number is assigned to each scenario to facilitate referring to them throughout this manuscript.

Designing a substantially higher number of models for case 4 is a voluntary decision and corresponds to an industry reality since most of the sub-assemblies include at least one $\overline{\text{SOR}}$ component (i.e. elbow, tee, valve, and reducer-eccentric). The probability of REF and ADD to be a $\overline{\text{SOR}}$ is thus higher and that is why case 4 includes more scenarios.

4.1 Impact of MRM on Reducing the Spatial Complexity and Number of Components

The effectiveness of the method is first evaluated in terms of the spatial complexity reduction and the decrease in the number of components. To this end, the concept of Level of Complexity (LOC) is introduced. LOC is a value that attempts to gather the spatial complexity and the number of components together. LOC is a driver of 3D scanning requirements. Each component is assigned a LOC detailed in Table 2.

For each component, key-points are established, and their associated unit vector is calculated. The number of unaligned vectors within a component provides the LOC for

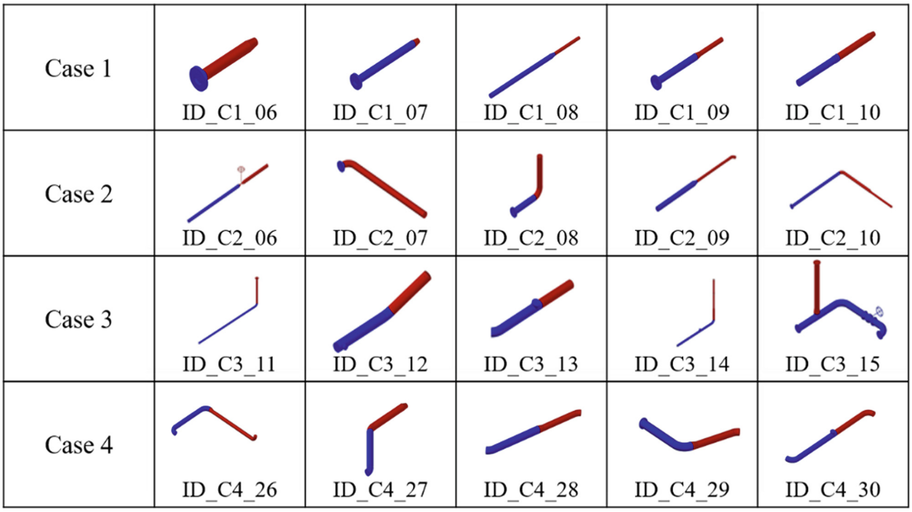


Fig. 7. 3D models of five scenario examples per case: components in blue belong to REF and those is red to ADD. (To access the full database, contact the corresponding author) (Color figure online)

Table 2. LOC value for each type of piping component

Type of component	Level of complexity
Pipe	1
Flange	1
Elbow	2
Reducer-concentric	1
Reducer-eccentric	2
Tee	2
Valve	2
Coupling	1
Cap	1

that component. Once each component has been assigned a value, the LOC of the full assembly can be calculated using Eq. 5.

$$\text{LOC}_{\text{assembly}} = \sum \text{LOC}_{\text{components}} \tag{5}$$

Consequently, the LOC handles at the same time the spatial complexity of the assembly and the number of components. The LOC values before and after applying the MRM is calculated for all of the designed scenarios and tabulated in Fig. 8.

In the four cases, all the MRMs have a reduced LOC. The reduction of the Level of Complexity is important when a $\overline{\text{SOR}}$ component is found as the first component of REF and/or ADD. The more components the initial assembly has, the more likely the

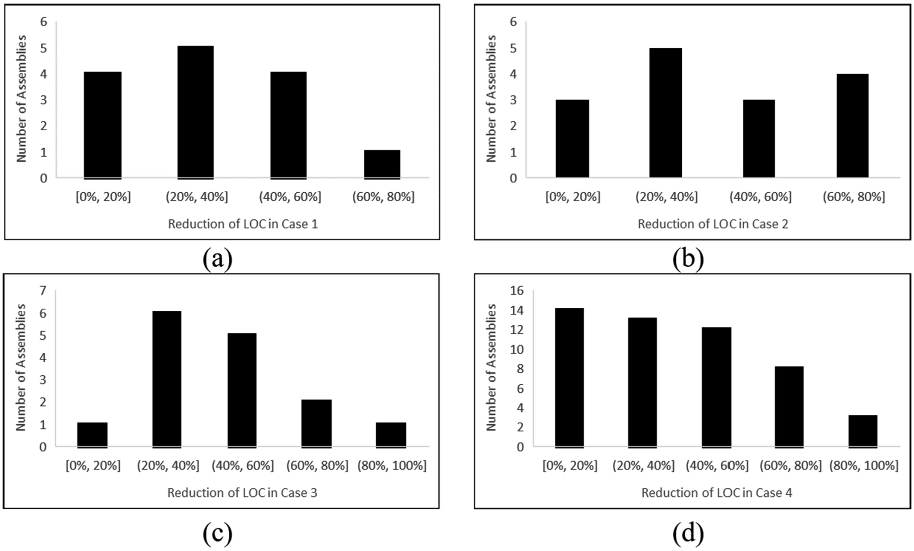


Fig. 8. LOC reduction between initial 3D models and after applying MRM for (a) Case 1, (b) Case 2, (c) Case 3 and (d) Case 4

MRM becomes useful. Thus, as shown in Table 3, the mean reduction of LOC for case 2, 3 and 4 is higher than case 1. This is due to the fact that all the components in case 1 are SOR and as explained earlier, there are limited assemblies that have this quality. Case 1, 2 and 3, have at least REF or ADD being SOR. As a result, the MRM can be reduced to the only two components being welded independently of the other components. In case 4, the reduction can be very important if the first two components are $\overline{\text{SOR}}$, but can also be very small when the first $\overline{\text{SOR}}$ components are far from the weld.

Table 3. Average LOC reduction for each case

Case number	Case 1	Case 2	Case 3	Case 4
Mean reduction of LOC	31%	40%	47%	41%

For example, in ID_C4_26, REF is composed of an elbow, a pipe and an elbow and ADD is a reducer-eccentric, a pipe and an elbow. The weld is between the elbow and the reducer-eccentric. Thus, the MRM is only composed of the elbow and the reducer-eccentric. The result of LOCs for ID_C4_26 is detailed in Table 4.

Table 4. Result of level of complexity for ID_C4_26

$\text{LOC}_{\text{InitialModel}}$	LOC_{MRM}	Reduction of LOC
10	4	60%

The reduced LOC is important since it directly correlates with the time required for inspection of an assembly. Furthermore, the lower the LOC is, the more accurate the acquisition result will be.

4.2 Impact of MRM on Collecting as-Built Information Using Mobile Scanners

As explained earlier, MRM is designed to be employed in a step-wise assembly control system for construction workers. To that end, it is important to measure its performance with respect to limitations of consumer grade scanning devices.

One of the major challenges of real-time mobile scanners are their limited scanning range and accuracy (Fig. 9). While being cost effective and easy to use, the accuracy of these type of sensors decreases as objects are placed further away. Consequently, it will be significantly beneficial to reduce the length of objects that need to be scanned. Figure 9 shows an example mobile scanner (maximum accuracy 0.5 cm for objects in less than 1-m distance) and its range limitation (scanning range 4 × 4 × 4 (m)). As shown, the maximum length that is feasible to obtain as-built information is 6.92 (m).

By defining principal length as the shortest line connecting the two ends of the assembly, a performance measure can then be defined (Eq. 6) for MRM with respect to components' length.

$$U_{MRM} = \frac{(K_1 - K_2)}{K_1} \times 100, \tag{6}$$

where K_1 denotes to the principal length of the assembly before applying MRM and K_2 denotes to the principal length of the component after applying MRM.

To more accurately quantify the effectiveness of MRM on addressing the above challenges Monte Carlo simulations were performed on the created database. It was assumed that the only contributor to assemblies' length are pipes. In order to simulate variability of pipes' length, expert's advice was sought and the assemblies were divided into 3 groups based on their length (Table 5).

The following results are based on 20 Monte Carlo simulations, each simulating 1000 events in each size category. To calculate U_{MRM} for each assembly, the principal lengths' were then calculated and recorded.

Each point in Fig. 10 corresponds to an assembly existing in the database and has 2 values associated with it, K_1 and K_2. Points in blue color show the K_1 and K_2 values for scenarios simulated for small assemblies, red points correspond to medium size assemblies, and finally, green points belong to assemblies made up with large pipes. As shown, the line $U_{MRM} = 0$, divides the space into 2 areas: A_4 and A_1, A_2 and A_3. As explained earlier, in the worst-case scenario, applying MRM will not reduce the LOC, and thus the length of the principal components will not change ($K_1 = K_2$). As anticipated, no point exists in the A_4 area. The remainder of space is divided into 3 areas by drawing the lines L_1 and L_2. Points in the A_3 area correspond to scenarios were the principal length of the assembly exceeds the scanner's range and thus obtaining as-built information is infeasible. These assemblies are too large to be scanned and employing MRM will have no utility. Points in the A_2 correspond to

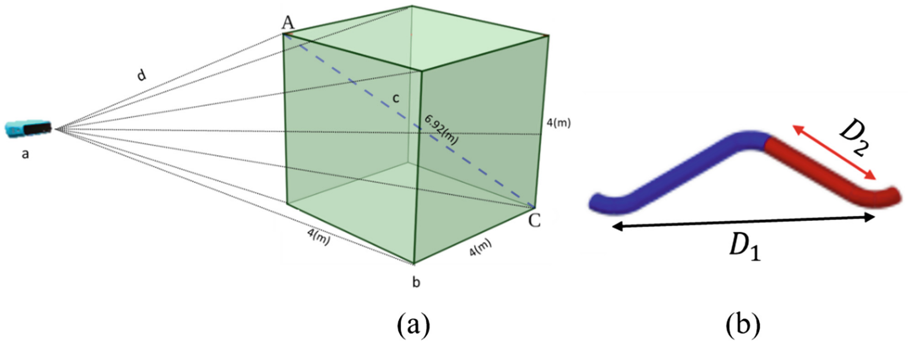


Fig. 9. (a) Scanning range and maximum feasible length for scanning. a: acquisition device, b: only objects within the projected cube can be scanned, c: maximum allowable length for an object to be scanned, assemblies that have a principal length of more than 6.92 (m) cannot be scanned and controlled. d: infrared beams projected to capture the scene. (b) Principal length before and after applying MRM. D_1 is the principal length of the assembly before applying MRM. D_2 is the principal length of the assembly after applying MRM.

assemblies that can only be scanned if MRM is applied. In other words, this area shows the assemblies whose as-built information could not be obtained without the help of MRM. Finally, points in the A_1 area can be scanned whether MRM is applied or not. However, applying MRM on some of the assemblies can reduce the principal length in the assembly and thus increase the accuracy of the obtained information. The diagonal line drawings in this area indicate how much the principal components have been reduced. For example, if a point lies between lines U_{MRM} = 40% and U_{MRM} = 60%, it has had a reduction of more than 40% and less than 60% in its principal length which directly correlates with the accuracy of the obtained as-built information. Table 6 shows the distribution of points within A_1, A_2 and A_3.

Table 5. Categorization of the assemblies based on average length of pipes in the assemblies and the intended application. The length of pipes in each category was assumed to follow a normal distribution.

Attributes	Small	Medium	Large
Mean Length (M)	1.5	3.5	5
Standard deviation	0.5	1	2
Application	Mostly residential and non-industrial	Nuclear and water supply	Oil and gas, large industrial projects, and pipe lines

Furthermore, in order to investigate the application and the performance of MRM within the three categories, Fig. 11 was prepared.

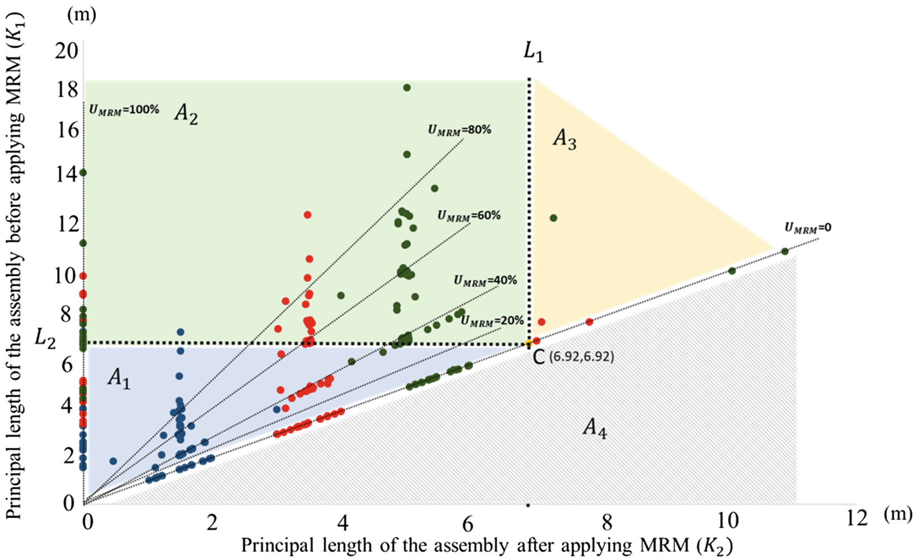


Fig. 10. Monte Carlo simulation results using the assemblies from the database with three size categories. (Color figure online)

Table 6. Distribution of points in the simulation and MRM's utility in each group.

Area	Ratio	MRM's Utility
A_1	61.0%	Increased acquisition accuracy
A_2	36.9%	Only feasible to inspect with MRM
A_3	2.1%	No utility

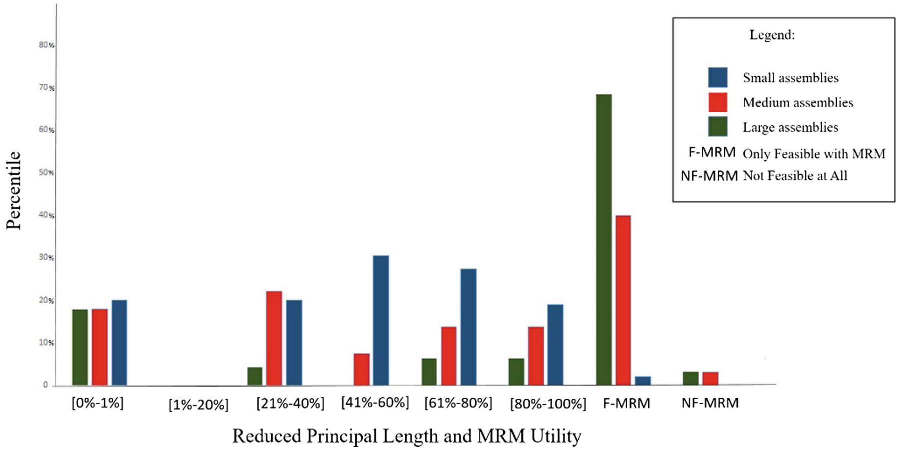


Fig. 11. Comparing MRM's utility within the three size groups

As can be seen, regardless of the assembly's size, MRM will not be useful for the assemblies that have a specific geometry (U_{MRM} = 0%). For instance, ID_C1_07, the initial model is composed of a flange, a pipe and a reducer-concentric. Applying MRM only removes the flange, which doesn't impact the principal length. Furthermore, the results show that applying MRM has the most impact on accuracy for assemblies with small dimensions. Finally, more than 70% of the assemblies are infeasible to obtain as-built information in large assemblies; this shows that applying MRM is critical for any step-wise assembly process control for large assemblies.

5 Conclusion and Future Work

Due to higher requirements of accuracy and economical consequences of any defect, controlling the quality of industrial assemblies in fab shops has become crucial. Manual tools are usually utilized to measure the dimensions and check the alignments. To reduce the number of tasks performed manually, 3D vision technologies have become common practice. However, these methods are often applied at the end of the fabrication process, because they can be cumbersome, time consuming and require engineering knowledge. The proposed method is necessary for any stepwise quality control process. By using the Minimum Required Model on a mobile scanner, workers can perform the inspection every time components are assembled.

The effectiveness of the Minimum Required Model in reducing the spatial complexity of the original assembly and in facilitating the utilization of consumer grade scanning devices was shown. The methodology has been experimented with on multiple industrial assembly models with different configurations and dimensions (length, diameter). Industrial applications utilizing consumer grade scanners could emerge by applying the MRM.

When designing the 3D model database, the objective was to make sure that the assemblies are realistic and simulate real scenarios encountered in the industry. The value provided is based on the designed models in this paper and may change with different designs. However, the provided framework to obtain the MRM is universal and can be applied to any assembly. Based on the designed database, the MRM's impact has been demonstrated to be more important on large and spatially complex assemblies. Most of the emerging scan vs model applications can thus use the MRM to positive effect.

In order to truly evaluate the effectiveness of the method, future work will be conducted and will consist in applying the proposed methodology for each step of the construction assembly fabrication. Following workers steps of assembling will enable gathering the real steps of fabrication and thus to apply the methodology for each step of the evaluated assembly. Several industrial assemblies will be investigated under the proposed measurements of effectiveness, and the utility of the MRM will be measured based on the complete assembly and not just on a single step of fabrication.

Other potentials of the MRM should be analyzed such as the cost savings. The authors believe that performing the stepwise quality control using the MRM with a consumer grade scanning device can substantially reduce rework and measurement time. Also, the use of such a technique would certainly reduce the number of incompliances detected at the end of the fabrication which could either be caused by a worker's mistake or by the propagation of tolerances along the assembly. Therefore, quantifying the savings of the described methodology in the overall construction process of industrial assemblies will be done as future work.

Acknowledgments. The authors would like to thank Stacey Jensen Rose and Scott Waters from Aecon, Canada, for their invaluable insights and feedback on this work. This research is partially funded by Natural Science and Engineering Research Council (NSERC) Canada, and Aecon Industrial West (AIW), Sherwood Park, Canada.

References

1. Jaillon, L., Poon, C.S.: Life cycle design and prefabrication in buildings: a review and case studies in Hong Kong. Autom. Constr. **39**, 195–202 (2014)
2. Jaillon, L., Poon, C.S.: Sustainable construction aspects of using prefabrication in dense urban environment: a Hong Kong case study. Constr. Manag. Econ. **26**(9), 953–966 (2008)
3. Gnanavel, C., Saravanan, R., Chandrasekaran, M., Jayakanth, J.J.: Improvement of productivity in TIG welding plant by equipment design in orbit. IOP Conf. Ser. Mater. Sci. Eng. **183**(1) (2017). 012020
4. Fatal occupational injuries by selected characteristics, 2003–2014 (2003)
5. Pan, W., Gibb, A.G., Dainty, A.R.: Strategies for integrating the use of off-site production technologies in house building. J. Constr. Eng. Manag. **138**(11), 1331–1340 (2012)
6. Goodrum, P.M., Miller, J., Sweany, J., Alruwaythi, O.: Influence of the format of engineering information and spatial cognition on craft-worker performance. J. Constr. Eng. Manag. **142**(9) (2016). 0401604
7. Safa, M., Shahi, A., Nahangi, M., Haas, C., Noori, H.: Automating measurement process to improve quality management for piping fabrication. In: Structures, vol. 3, pp. 71–80. Elsevier (2015)
8. Omar, T., Nehdi, M.L.: Data acquisition technologies for construction progress tracking. Autom. Constr. **70**, 143–155 (2016)
9. Sharif, M.M., Nahangi, M., Haas, C., West, J., Ibrahim, M.: A preliminary investigation of the applicability of portable sensors for fabrication and installation control of industrial assemblies. In: Canadian Society for Civil Engineering (2016)
10. US Census Bureau, Construction Spending. https://www.census.gov/construction/c30/c30index.html
11. Patterson, L., Ledbetter, W.B.: The cost of quality: a management tool. In: Excellence in the Constructed Project, pp. 100–105. ASCE (1989)
12. Akinci, B., Boukamp, F., Gordon, C., Huber, D., Lyons, C., Park, K.: A formalism for utilization of sensor systems and integrated project models for active construction quality control. Autom. Constr. **15**(2), 124–138 (2006)
13. Leite, F., Cho, Y., Behzadan, A.H., Lee, S., Choe, S., Fang, Y., Hwang, S.: Visualization, information modeling, and simulation: grand challenges in the construction industry. J. Comput. Civ. Eng. **30**(6) (2016). 04016035
14. Holzer, D.: Are you talking to me? Why BIM alone is not the answer (2007)
15. Chen, L., Luo, H.: A BIM-based construction quality management model and its applications. Autom. Constr. **46**, 64–73 (2014)

16. Farnsworth, C.B., Beveridge, S., Miller, K.R., Christofferson, J.P.: Application, advantages, and methods associated with using BIM in commercial construction. Int. J. Constr. Educ. Res. **11**(3), 218–236 (2015)
17. Garcia, M.A., Llanos, D.R., De Prada, C.: A configurable ACSL-based interface generator for simulated systems. Simulation **73**(4), 206–212 (1999)
18. Lee, J., Lee, K., Nam, B., Wu, Y.: IoT Platform-based iAR: a prototype for plant O&M applications. In: IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), pp. 149–150. IEEE (2016)
19. Fung, M., Fung, A., Fung, R.H., Fung, P.K.: The advantage of using a file data transfer method in a plant design. In: ASME 2014 Pressure Vessels and Piping Conference, p. V003T03A021. American Society of Mechanical Engineers (2014)
20. Newman, T.S., Jain, A.K.: A survey of automated visual inspection. Comput. Vis. Image Underst. **61**(2), 231–262 (1995)
21. Bosché, F., Ahmed, M., Turkan, Y., Haas, C.T., Haas, R.: The value of integrating scan-to-BIM and scan-vs-BIM techniques for construction monitoring using laser scanning and BIM: the case of cylindrical MEP components. Autom. Constr. **49**, 201–213 (2015)
22. Nahangi, M., Czerniawski, T., Haas, C.T., Walbridge, S., West, J.: Parallel systems and structural frames realignment planning and actuation strategy. J. Comput. Civ. Eng. **30**(4) (2015). 04015067
23. Sharif, M.M., Nahangi, M., Haas, C., West, J.: Automated model-based finding of 3D objects in cluttered construction point cloud models. Comput. Aided Civ. Infrastruct. Eng. **32**(11), 893–908 (2017)
24. Wang, J., Sun, W., Shou, W., Wang, X., Wu, C., Chong, H.Y., Sun, C.: Integrating BIM and LiDAR for real-time construction quality control. J. Intell. Robot. Syst. **79**(3–4), 417 (2015)
25. Tang, P., Akinci, B.: Automatic execution of workflows on laser-scanned data for extracting bridge surveying goals. Adv. Eng. Inform. **26**(4), 889–903 (2012)
26. Bosché, F.: Automated recognition of 3D CAD model objects in laser scans and calculation of as-built dimensions for dimensional compliance control in construction. Adv. Eng. Inform. **24**(1), 107–118 (2010)
27. Malamas, E.N., Petrakis, E.G., Zervakis, M., Petit, L., Legat, J.D.: A survey on industrial vision systems, applications and tools. Image Vis. Comput. **21**(2), 171–188 (2003)
28. Bosché, F., Guillemet, A., Turkan, Y., Haas, C.T., Haas, R.: Tracking the built status of MEP works: assessing the value of a scan-vs-BIM system. J. Comput. Civ. Eng. **28**(4) (2013). 05014004
29. Bosché, F., Haas, C.T., Akinci, B.: Automated recognition of 3D CAD objects in site laser scans for project 3D status visualization and performance control. J. Comput. Civ. Eng. **23** (6), 311–318 (2009)
30. Golparvar-Fard, M., Peña-Mora, F., Savarese, S.: Integrated sequential as-built and as-planned representation with D 4 AR tools in support of decision-making tasks in the AEC/FM industry. J. Constr. Eng. Manag. **137**(12), 1099–1116 (2011)
31. Tang, P., Anil, E.B., Akinci, B., Huber, D.: Efficient and effective quality assessment of as-is building information models and 3D laser-scanned data. Comput. Civ. Eng. **2011**, 486–493 (2011)
32. Turkan, Y., Bosché, F., Haas, C.T., Haas, R.: Automated progress tracking using 4D schedule and 3D sensing technologies. Autom. Constr. **22**, 414–421 (2012)
33. Anil, E.B., Tang, P., Akinci, B., Huber, D.: Deviation analysis method for the assessment of the quality of the as-is building information models generated from point cloud data. Autom. Constr. **35**, 507–516 (2013)

34. Brilakis, I., Lourakis, M., Sacks, R., Savarese, S., Christodoulou, S., Teizer, J., Makhmalbaf, A.: Toward automated generation of parametric BIMs based on hybrid video and laser scanning data. Adv. Eng. Inform. **24**(4), 456–465 (2010)
35. Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A.: Automatic reconstruction of as-built building information models from laser-scanned point clouds: a review of related techniques. Autom. Constr. **19**(7), 829–843 (2010)
36. Bhatla, A., Choe, S.Y., Fierro, O., Leite, F.: Evaluation of accuracy of as-built 3D modeling from photos taken by handheld digital cameras. Autom. Constr. **28**, 116–127 (2012)
37. Wang, X., Truijens, M., Hou, L., Wang, Y., Zhou, Y.: Integrating augmented reality with building information modeling: onsite construction process controlling for liquefied natural gas industry. Autom. Constr. **40**, 96–105 (2014)
38. Hou, L., Wang, X.: Experimental framework for evaluating cognitive workload of using AR system for general assembly task. In: Proceedings of the 28th International Symposium on Automation and Robotics in Construction (2011)
39. Makris, S., Pintzos, G., Rentzos, L., Chryssolouris, G.: Assembly support using AR technology based on automatic sequence generation. CIRP Ann. Manuf. Technol. **62**(1), 9–12 (2013)
40. Wang, X., Ong, S.K., Nee, A.Y.C.: Multi-modal augmented-reality assembly guidance based on bare-hand interface. Adv. Eng. Inform. **30**(3), 406–421 (2016)
41. Aiteanu, D., Hillers, B., Graser, A.: A step forward in manual welding: demonstration of augmented reality helmet. In: Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 309–310. IEEE (2003)
42. Schwerdtfeger, B., Pustka, D., Hofhauser, A., Klinker, G.: Using laser projectors for augmented reality. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pp. 134–137. ACM (2008)
43. Molleda, J., Usamentiaga, R., García, D.F., Bulnes, F.G., Espina, A., Dieye, B., Smith, L.N.: An improved 3D imaging system for dimensional quality inspection of rolled products in the metal industry. Comput. Ind. **64**(9), 1186–1200 (2013)
44. Siebert, S., Teizer, J.: Mobile 3D mapping for surveying earthwork projects using an Unmanned Aerial Vehicle (UAV) system. Autom. Constr. **41**, 1–14 (2014)
45. Irizarry, J., Costa, D.B.: Exploratory study of potential applications of unmanned aerial systems for construction management tasks. J. Manag. Eng. **32**(3) (2016). 05016001
46. Fathi, H., Dai, F., Lourakis, M.: Automated as-built 3D reconstruction of civil infrastructure using computer vision: achievements, opportunities, and challenges. Adv. Eng. Inform. **29** (2), 149–161 (2015)
47. Golparvar-Fard, M., Bohn, J., Teizer, J., Savarese, S., Peña-Mora, F.: Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. Autom. Constr. **20**(8), 1143–1155 (2011)
48. Kim, C., Son, H., Kim, H., Han, S.H.: Applicability of flash laser distance and ranging to three-dimensional spatial information acquisition and modeling on a construction site. Can. J. Civ. Eng. **35**(11), 1331–1341 (2008)
49. Chien, K.F., Wu, Z.H., Huang, S.C.: Identifying and assessing critical risk factors for BIM projects: empirical study. Automation in Construction **45**, 1–15 (2014)
50. Jeffery, G.B.: On the steady rotation of a solid of revolution in a viscous fluid. Proc. Lond. Math. Soc. **2**(1), 327–338 (1915)

Quantitative Analysis of Close Call Events

Olga Golovina, Manuel Perschewski, Jochen Teizer(✉),
and Markus König

Ruhr-University Bochum, Bochum, Germany
{olga.golovina,manuel.perschewski,
jochen.teizer}@rub.de, koenig@inf.bi.rub.de

Abstract. Construction safety is a big problem according to official statistics. In many of the developed countries about 15–25% of all fatal construction workplace accidents relate to too close proximity of pedestrian workers to construction equipment or hazardous materials. Extracting knowledge from data to near hits (aka. close calls) might warrant better understanding on the root causes that lead to such incidents and eliminate them. While a close call is a subtle event where workers are in close proximity to a hazard, its frequency depends – amongst other factors – on poor site layout, a worker's willingness to take risks, limited safety education, and pure coincidence. Some pioneering organizations have recognized the potential on gathering and analyzing leading indicator data on close calls. However, mostly manual approaches are infrequently performed, subjective due to situational assessment, imprecise in level of detail, and importantly, reactive or inconsistent in effective or timely follow-ups by management. While existing predictive analytics research targets change at strategic levels in the hierarchy of organizations, personalized feedback to strengthen an individual worker's hazard recognition and avoidance skill set is yet missing. This study tackles the bottom of Heinrich's safety pyramid by providing an in-depth quantitative analysis of close calls. Modern positioning technology records the trajectory data of personnel, equipment, and materials. Computational algorithms then automatically generate previously unavailable details to close call events. The derived information is embedded in simplified geometric information models that users on a construction site can retrieve, easily understand, and adapt in existing preventative hazard recognition and control processes. Results from scientific and field experiments demonstrate that the developed system works successfully under the constraints of currently available positioning technology.

Keywords: Construction safety · Close calls · Predictive analytics

1 Introduction

Better understanding the root causes that lead to an accident is important to protect construction personnel from similar mishaps in the future. Unfortunately, most of the current accident investigation methods focus on supplying valuable information after the fact, once a person has been injured or killed.

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 359–384, 2018.
https://doi.org/10.1007/978-3-319-91635-4_19

Accident investigation reports, as explained in [1], are often brief and only a few pages long [2]. Fatality assessment and control evaluation (FACE) reports are one example of a practiced method of an investigation [3]. They typically contain factual information, for example: a description of what happened, the actual results of the event, the persons involved, the equipment or material involved, the activities preceding and during the event, the date, time and place of the event, any emergency actions taken, some pictures of the event situation, and the immediate remedial actions taken. They may also include additional information: risk classification, determination of potential consequences, cause analysis, direct causes, basic causes, management system factors, and importantly, remedial actions which include the assignment of responsibilities for adequate follow-ups.

The professional, who conducts the written investigation, usually enters the reporting process in three ways: (a) designs the report forms and keeps them current for the organization, (b) analyzes the data for trends and implications, and (c) measures of the quality of the report (typically with a manual scoring sheet to enable continuous improvement of the reporting and follow-up processes). Conducted in such manner, the professional can promote thorough investigations and quality reports which enable full control by management later.

While the contributions of this study do not substitute any of the existing investigation approaches that are in place, it tackles the topic more pro-actively. In the ideal case, the proposed method should support existing processes with new information that has not been made available before. As [4] has previously outlined, construction safety has to happen at the right-time. Thanks to emerging technology, detailed information on close calls can be recorded and analyzed near real-time. The generated information can then be used for immediate mitigation or even predictive analysis.

This paper first reviews the existing research on close calls in construction, then explains the proposed algorithm for quantitative analysis of close call events in construction safety. Scientific verification through simulation and validation using real field experiments follow. The results demonstrate the functionality of the developed algorithm and software user interfaces. A discussion and an outlook for future research conclude the paper.

2 Background

A vast body of knowledge exists on close calls within the construction industry and outside of it. This evidence in the published literature is not repeated, instead this review focuses on a comparison of manual and automated data collection methods that are suitable for close call measurements only.

2.1 Close Calls

Several researches in construction describe a close call as an event that resulted almost in an accident. Too close proximity between a pedestrian worker and a known hazard is one of such issues. However, there is no research that provides a scientific definition of the exact characteristics of a close call [5]. According to [6], a close call can be part of a sequence of events that result in minor or major accidents. Therefore, close calls should be recorded and followed-up with a close call reporting program. Such program measure safety performance and reduce the probability of accidents. However, the success of close call reporting crucially depends on the participation of persons to report near-misses, which can lead to inconsistent or false results [7].

2.2 History on Reporting and Analyzing Close Calls

Heinrich's safety pyramid (aka. accident triangle) provides an early example for separating close calls (called therein *near misses*) from actual accidents. Visualization the difference between accidents (e.g., fatality and injury) and incidents (e.g. at-risk behavior and close calls) in a graph strengthens the case for the higher occurrence of close calls relative to the number of fatal accidents or injuries. Interestingly to note, the original data to generate the safety pyramid came from a manual analysis of 75,000 injury and illness reports [8].

Fast forward and decades later, the results from a survey by [9] suggest that employees from companies with high health and safety ratings perceive their own safety, zero harm, and continuous improvement in health and safety as very important. In this study, construction hazard identification, including close call reporting, ranked 10th out of 38 which shows the general acceptance of such a system. [10, 11] then discussed strengths and weaknesses for a qualitative (matrix) and quantitative (index) near-miss management system. They focused on how close call reporting and filtering could be implemented to minimize both missed near-miss reports and unnecessary reports. Their design consists of four separate phases: Event identification and reporting, event assessment, prevention measure application and follow-up actions. Among other noteworthy research that followed, [12] for example, established a database consisting of feature vectors (values that represent information on an incident) for close calls, filled with data from common written incident-reports, viewing close calls as events which lead to an accident.

Today, under often self-motivated initiatives for establishing leading indicators for safety, pioneering owner and contractor organizations highly encourage the (voluntary) reporting and analysis of close calls by everyone involved in a project. Databases with restricted access exist where close calls are entered manually or via guided user interfaces (GUI) on mobile devices. These examples from modern construction sites demonstrate the advancements that have been made for reporting and investigating incidents recently. In brief, the reasons for this change can be summarized twofold: (a) driving organizational change in safety culture by rethinking existing and establishing new processes and (b) taking advantage of sophisticated technologies to record and analyze real data.

2.3 Related Examples Using Technology

The most closely related previous study was performed by [13]. It describes a method called Proximity Hazard Indicator (PHI). PHI successfully detects spatial-temporal (proximity) conflicts between workers and construction equipment using real-time location sensing (RTLS). Other researchers, for example [14] used a real-time location and a virtual construction simulation system to test the performance on safety behavior. [15, 16] also demonstrated in applications of virtual reality (VR) when pedestrian workers are getting too close to heavy construction equipment. Several more research groups identified that real data, including construction site layout and building geometry from BIM and trajectories of workers and equipment from RTLS, would make their VR-scenarios more realistic.

[17] developed a tracking system of near-miss accidents on construction sites to aid in the research of accident prevention on construction sites. Proposed system uses Zigbee radio frequency identification (RFID) to identify resources and store specific information (ladders' last time of inspection), ultrasound tags to track the location and sensors to measure environmental information (brightness, noise, weather). Though this method allows the detection of more subtle and complex accident precursors, it focuses little on human-machine interaction.

[18] first integrated Ultra-Wideband (UWB), a wireless location tracking technology, in the practical training workflow of union ironworkers. They collected data for post-reasoning lagging safety and productivity indicators. One suggestion of their work is to improve the worker education and training by personalizing feedback. They envision on using near real-time analysis of actual training data. Importantly, they also conclude that any technology that assists in such a case, must be wisely selected. UWB, for example, requires a rather larger investment and set-up of sensing infrastructure. UWB, like many other remote sensing approaches, may also be limited by its signal strength (causing at minimum location measurement errors).

[19] introduced CHASTE (Construction Hazard Assessment with Spatial and Temporal Exposure) which assigns estimated risk levels to specific tasks to compute risk-levels of scenarios. Although they might have years of experience the CJHA requires employees to manually evaluate the risk level of specific construction tasks, leading to potential errors or inconsistency. If the construction plans change, new tasks might have to get evaluated. [20–23] presented examples for utilizing construction safety knowledge to improve Job Hazard Analysis (JHA). While their approach requires human input by experts.

[24, 25] pursued an alternative method of detecting hazards for outdoor work environments. In their respective works, Global Navigation Satellite System (GNSS) data loggers record the resources' location (work crews and equipment, respectively) and visualized the associated close call risks using heatmaps. Safe work station planning based on real-time resource location tracking and site layout geometry data becomes possible. Both studies refer to [26] who performed an in-depth evaluation on commercially-available GNSS data loggers.

2.4 Remaining Problems

Practiced close call reporting and analysis rely on manual data gathering efforts. Using only manual reports as a source of information causes several disadvantages. Some of the issues presented next help explain the problem:

1. *Size of the problem:* The number of reported close calls is probably smaller than the real number (i.e., workers may not report close calls fearing retaliation or a drop in productivity).
2. *Standardization:* Accident investigation reports vary by country and are kept general to inform the entire organization and sometimes even the industry (i.e., an open-access benchmark based on high quality, near real-time data and available to every construction site and personnel is missing).
3. *Data availability and processing:* Processes depending on manual data lack the level of detail (i.e., unlike in the airline industry for the past decades or unmanned vehicles just recently, trajectories of construction equipment are often neither recorded nor analyzed).
4. *Collaborative planning:* Though Building Information Modeling (BIM) offers the construction industry a method to plan, build, and operate infrastructure or buildings, standardized tools for construction safety (and health), site layout or work station planning are missing (i.e., most projects perform modeling efforts with BIM manually at low detail and only on an as-needs basis).
5. *Safety culture change for labor and management:* Since close call reports may include sensitive information to an incident [27], person(s) reporting them might impact labor-management (i.e., workforce vs. supervisor, management) relations and organizational fairness.

3 Objectives and Scope Definition

Using only manual approaches to gather information about close calls is not practicable as these can be subtle and frequent events and the assessment might vary depending on the observer. Therefore, the novelty of the proposed approach is to supplement additional information to the close call recording, analysis, and follow-up processes. Since human-machine interactions are one of the more serious problems in the construction industry [28], the study's specific focus is on investigating human-equipment and human-hazardous material incidents. By doing so the traditional close call reporting and follow-up process is changed (see Fig. 1).

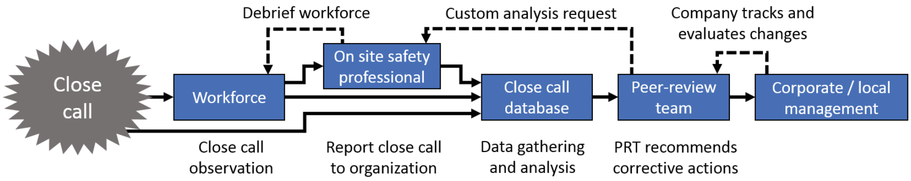

Debrief workforce
Custom analysis request
Company tracks and evaluates changes
Close call
Workforce
On site safety professional
Close call database
Peer-review team
Corporate / local management
Close call observation
Report close call to organization
Data gathering and analysis
PRT recommends corrective actions

Fig. 1. Close call reporting, analysis, and personalized feedback process

Close calls, as introduced earlier, are typically reported when a person witnesses or participates in an event which compromises or threatens to compromise the health or safety of a person or the environment. If necessary, the person may conduct first efforts to prevent an accident or further incident. The person notifies their supervisor or safety coordinator on site directly or using a close call reporting application on a mobile device (i.e., if permitted on site: smartphones or tablets). Some organizations offer close-call reports through a neutral third party to remove sensitive information. Once the case reaches these knowledgeable persons, they contact, if not happened immediately, the corresponding safety professional within the organization. By sharing at least some general information about the event, a problem-solving peer-review team consisting of workforce (who are trained in operational skills), safety professionals (who are trained in root-cause analysis), and management (who are trained in continuous-process improvement) will heighten the awareness in their own organization. Various means exist to learn more about the risks and how to mitigate them, for example, dedicated close call review meetings, department safety meetings, one-on-one with workforce or supervisors, or involving a neutral third party. The team, while protecting employees from blame [29], finally recommends corrective actions. Well-working close call reporting processes ensure timely feedback to the person who reported the incident in the first place.

4 Proposed Framework

The proposed framework takes advantage of remote sensing to automatically record the circumstances that lead to close calls. By attaching a precise real-time location sensing (RTLS) device on every resource (personnel, equipment, material that was a priori declared hazardous), their trajectory data will be analyzed to locate close calls and interfere further valuable information. The information generated therefore provides an elevated level of detail that has not been available so far. This way, measurement and evaluation of close calls during the actual construction phase becomes an active leading indicator which can result in an immediate improvement of safety performance [30].

A process was developed to automatically detect and analyze close call events between pedestrian workers, construction equipment, and other known hazards. These are named from now on *resources*. Further descriptive information to each individual resource involved in a *close call event* was made available, for example, its precise position and boundary information. Latter one defines a two-dimensional *protective envelope* (Fig. 2). For the reason of simplicity, all data presented in this study is kept to two-dimensional (2D or plan view) only. As a result, protective envelopes come in shapes of circles or polygons.

The initial development of the proposed data analysis tool for investigating the detail to close calls between resources involved the generation of a simulated data set in a fictional construction setting. The selection of simulated over realistic data permitted the verification and validation of the proposed method under ideal conditions. A building information model and trajectory information was assumed for pedestrian worker and equipment travel paths. From the resources' trajectories a first attempt was started to detect all proximity events and provide further insights to every event. The number of involved resources as well their parameters, i.e. the size of the protective envelope is called *safety distance*, were set in advance based on previous research

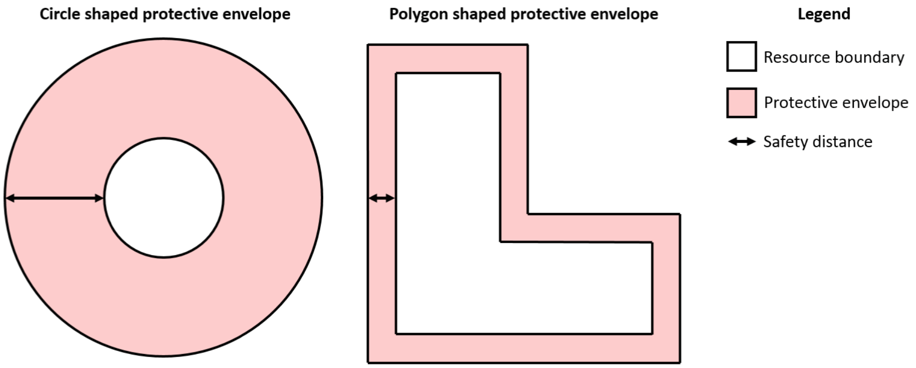

Circle shaped protective envelope
Polygon shaped protective envelope
Legend
Resource boundary
Protective envelope
Safety distance

Fig. 2. Examples of protective envelopes in plan view

findings by [31]. All values, except the trajectory information and building information model, remained unchanged in the evaluation.

As needed later in the experimental field validation, actual geometric data of the as-built conditions of the work environment were recorded using laser scanners and unmanned aerial vehicles (UAV) [32]. Their point cloud information was georeferenced and imported as simplified boundary objects in building information models [21]. The resource trajectory data from the outdoor work environments was recorded using remote sensing technology. Ultra Wideband (UWB) [31] and Global Navigation Satellite System (GNSS) [25] offered two suitable options to record such data. Field deployment of any of such real-time location sensing (RTLS) technology depends highly on the work environment that is under investigation. Business and technological factors, such as return on investment (ROI), signal propagation, size of measurement errors, hardware form factors, power consumption, ease of installation and maintenance, and many more items must be considered [31].

A workflow for fusing all data types and post processing generated descriptive analytics on each close call event (see Fig. 3). A test occurred first in a simulated scene. Details to each resource were discovered, for example, the course of close calls, individual resource- and hazard-statistics, a heatmap as well as comprehensive construction site safety statistics. All generated information is displayed on a layered Guided User Interface (GUI) (implemented in MATLAB®) which permits a user to assess construction safety from multiple view points and levels of detail. The GUI was designed based on industry expert input in a way to find intuitive answers to typical safety-performance-related questions:

- Which are the areas where pre-defined close calls occur frequently?
- Which workers or pieces of equipment are involved in a close call and are there particular differences in the safety performance among them?
- How does a worker react on entering a hazard zone and when might the worker recognize to be at risk and react upon detecting it?
- Which ways exist to leverage the newly generated information for continuous safety performance improvement, e.g. in safety education and training?

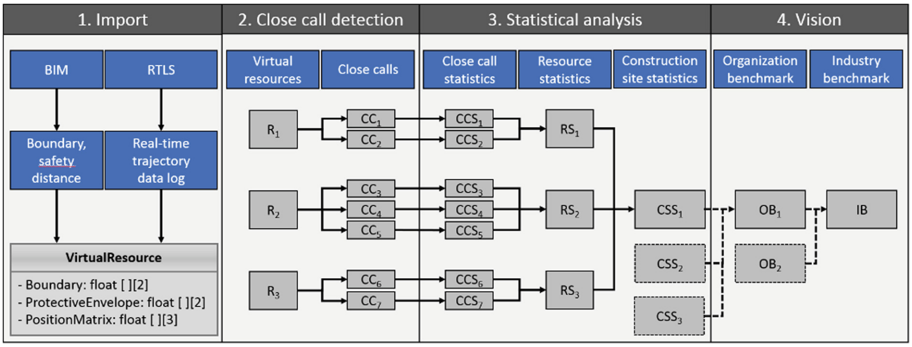


Fig. 3. Workflow for prototype development of data processing algorithm (dashed lines are part future of predictive close call data benchmarking)

5 Definitions

5.1 Construction Resource Data

Construction resources are physical objects and spaces that are required to finish a construction process. In this research, the term *construction resource* refers to (a) pedestrian workforce, (b) construction equipment, and (c) objects or structures of temporary or final state. The number of any of these resources in the scene under investigation can be one or many. They can also be static or dynamic in nature. Pedestrian workers as well as equipment are moving frequently, while temporary objects, such as scaffolds or hazardous materials like gas bottles, are mostly static in one position. Examples of static or as-built structures are elevator shafts or leading edges in high-rise construction.

Construction resource data is defined as a term to summarize boundary data from a building information modeling (BIM) and trajectory data from trajectory logging files. Microsoft EXCEL®-files served as the initial medium to transfer these information, since construction personnel is mostly familiar with this software package. The data for each resource is contained in a separate file.

5.2 Boundary Data Representing Resources

Boundary data represents a simplified version of the true shape of a resource in 2D space, typically derived from a building information model. Examples are presented for the different resource types. While a straight wall object is represented as a rectangle of the same length and width in 2D, workforce and equipment are simplified. The width of a human shoulder is approximately 0.6 m [31]. This value is rounded up to 1 m, representing the shape of a pedestrian worker as a circle. Slow speeds and rapid change in direction suits this representation of a worker well. In most application scenarios the simplified shape of equipment is a bounding box. A bounding box [33] encompasses all of its attachments inside. More complex objects are represented as a freeform using polygons. As explained earlier, boundary data contains a *safety distance* which extends the object boundary and creates a *protective envelope*.

5.3 Protective Envelopes

Unless specified otherwise by a user upfront, every resource boundary is surrounded by its own protective envelope (see Fig. 2). While the protective envelope is used to detect too close proximity events between resources, their size of the safety distance and shape are based on the following assumptions:

- *Pedestrian workforce:* A circle with a radius of 1.5 m is selected. This value is based on the average distance a human travels in one second, react, and come to a complete stop at such speed [31].
- *Construction equipment:* A protective envelope for equipment must be wisely chosen considering several of its operating parameters. These include, but are not limited to: operating speed, angle of operation, and articulation. While [34] has shown that multiple hazard zones for equipment are advisable to avoid a hit, generally a fixed value a user decides is added around the equipment's known bounding box.
- *Temporary object:* The size of its protective envelope is determined according to rules and regulations set by governments and local authorities [35]. The resulting shape is a resized version of the existing boundary.
- *As-built structure:* Structures which, once they are erected and remain on site, may also require protection. Guardrails, for example, preventing workforce or equipment from falling to lower levels typically have a protective envelopes associated to them. Their safe installation is also regulated by official regulations or company best practices [35].

5.4 Trajectory Data

Trajectory or position logging devices frequently store a resource's relative position and the current time, namely timestamps, inside a log-file [25, 26]. The logging frequency and additional logging information like battery status depend on the type of device. In this research, a frequency of one second is assumed to simplify the following calculations. When a log file is imported, its containing information is trimmed to a uniform trajectory matrix of the form:

$$T(R) = \begin{pmatrix} x_{start} & y_{start} & t_{start} \\ x_{start+1} & y_{start+1} & t_{start+1} \\ \vdots & \vdots & \vdots \\ x_{end+1} & y_{end+1} & t_{end+1} \\ x_{end} & y_{end} & t_{end} \end{pmatrix} \quad (1)$$

where t_{start}, t_{end} refer to the first and last logged timestamps and x and y to the location of the device. This matrix is referred to as trajectory data. To help with further definitions, a function which returns the position of *resource R* for a specific *timestamp t* is defined as:

$$P(R,t) = \begin{cases} (x_t, y_t),\ \mathit{if}\ t \in [t_{start}, t_{end}]; x_t, y_t, t, t_{start}, t_{end} \in T(R) \\ \mathit{undefined}, \mathit{else} \end{cases} \tag{2}$$

5.5 Close Call Event

Currently, there exists no common definition for close calls [5, 6]. A *close call*, as defined in this research, is a proximity event between one or several workers and a hazard, leading to an endangerment of the workers. A close call as it relates to a too close proximity event between two resources A and B is defined as an overlapping of their protective envelopes at positions $P(A,t)$ and $P(B,t)$. When using trajectory data, there are two possible approaches towards categorizing close call events: (a) to categorize every proximity event as a separate close call or (b) to conclude consecutive occurring proximity events to a single close call. Since a worker requires time to detect and react to endangerment and that multiple location data offers the possibility to conduct a more detailed analysis of the close call, the second approach is the more sensible choice for this research.

5.6 Close Call Event Buffering

For each proximity event, a buffer stores the event information for later processing. This information includes timestamp, position [m], velocity [m/s], orientation [°] and distance [m] towards the other resource and facing direction [°]. In the example shown in Fig. 4, a piece of equipment has been traversing too close to gas bottle.

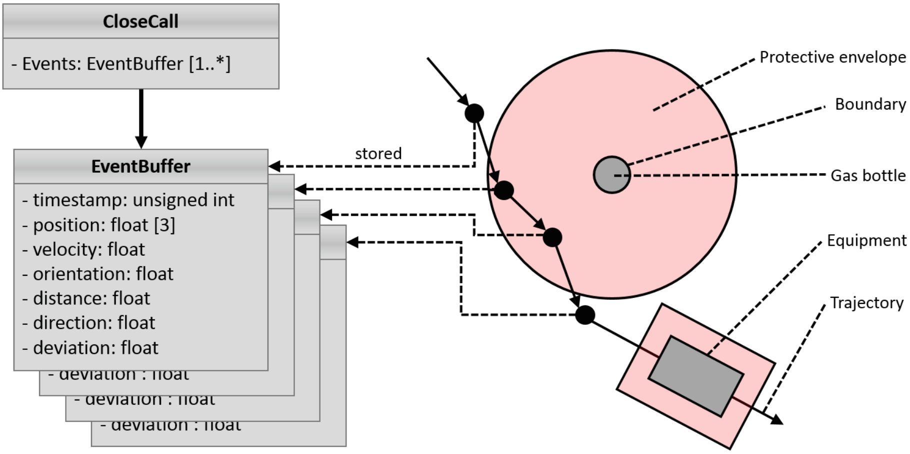


Fig. 4. EventBuffer class diagram

5.7 Close Call Analysis

For two resources A and B, a close call detection algorithm (1) analyses their trajectories and (2) checks for each *timestamp* $t \in T(A), T(B)$ if their protective envelopes overlap. If an overlap is found, a new close call gets created and a proximity event buffer is assigned to it. Every consecutive proximity creates a new event buffer which is added to the same close call. If no further overlap is detected, the close call is completed and the next proximity will create a new close call. Inside a completed close call, three event buffers will be marked for later processing:

- *Entry event:* First assigned event buffer.
- *Exit event:* Last assigned event buffer.
- *Closest event:* Event buffer where the distance between both resources is the smallest.

Additionally, the buffer events from the entry event to the closest distance event are summarized to the *entry path* and likewise the events from the closest distance event to the exit event are summarized to the *exit path*. As the trajectory data only consists of coordinates and timestamps, *velocity*, *facing direction*, *distance*, and *orientation* must be calculated separately.

5.8 Velocity

The close call algorithm has to compute a distinct velocity for each event buffer using only the resources' position data. As the trajectory logging frequency is assumed to be 1 Hz, the *velocity [m/s]* of a resource for *timestamp* t_i is equal to the 2D-Euclidean distance between $P(A, t_{i-1})$ and $P(A, t_i)$.

$$Vel(A, t_i) = \begin{cases} 0, t_{\text{start}} = t_i \\ Euclid(P(A, t_{i-1}), P(A, t_i)); t_{start} < t_i \leq t_{end} \\ undefined, else \end{cases} \tag{3}$$

5.9 Facing Direction

The direction towards which workers or vehicles are facing at a timestamp t_i is expressed as a normalized 2D-vector on the x-y-plane. Similar to the calculations for velocity, this vector can also be computed by using two position vectors. To be consistent, the direction will be calculated using $P(A, t_i)$ and $P(A, t_{i-1})$. Let *norm* be a function that returns the normalized version of a vector. Then the facing direction of a dynamic resource at *timestamp* t_i is defined as

$$Direction(A, t_i) = \begin{cases} norm(P(A, t_i) - P(A, t_{i-1})), t_{\text{start}} < t \leq t_{end} \\ Direction(t_{i+1}), t == t_{\text{start}} \\ undefined, else \end{cases} \tag{4}$$

5.10 Distance

For a *timestamp* t_i the distance between two resources is defined as the closest distance between their boundaries (see Fig. 5). The vector spanning this distance is described as the *boundary distance vector.* As these calculations are based on simple geometric operations, they are not discussed in greater detail.

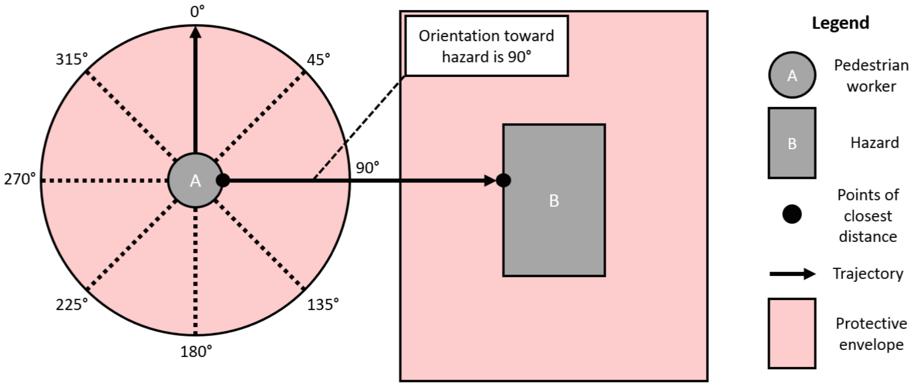


Fig. 5. Orientation

5.11 Orientation

The orientation value for an event buffer quantifies the position of the hazard relative to the facing direction of the resource. For this purpose, the resources' *facing direction vector* as well as the *boundary distance vector* will be utilized to compute an angle from 0° to 360°. The angle expresses by how many degrees a worker has to turn to the right to face the hazard directly (see Fig. 5).

6 Algorithm for Automated Close Call Data Processing

6.1 Trajectory Analysis

After storing all proximity event buffers, the close call analysis algorithm post-processes each close call to extract statistical information:

- *Duration:* The duration of the close call event in seconds. Under the assumption, that the logging frequency equals 1 Hz, the duration is equal to the number of event buffers.
- *Entry duration:* The time interval between entry event and closest event (including the closest event).
- *Exit duration:* Duration between closest event and exit event (excluding the closest event)
- *Hazard weights:* Values which indicate the severity of a close call. This includes a separate weight for orientation, velocity, distance, deviation, and duration.

Additionally, the *deviation* from an optimal direct path (see Fig. 6) is calculated. This direct path is assumed to be a path that leads directly from the entry position over the closest position to the exit position. It is calculated using the same number of steps as the real trajectory. The direct path positions are calculated by using a linear spacing algorithm between the entry position and closest position and between the closest position and exit position, respectively. In the following, the ratio between length of real path and length of direct path is described as the deviation of the close call. This value indicates how much the worker or vehicle has strayed from the optimal path during the close call event.

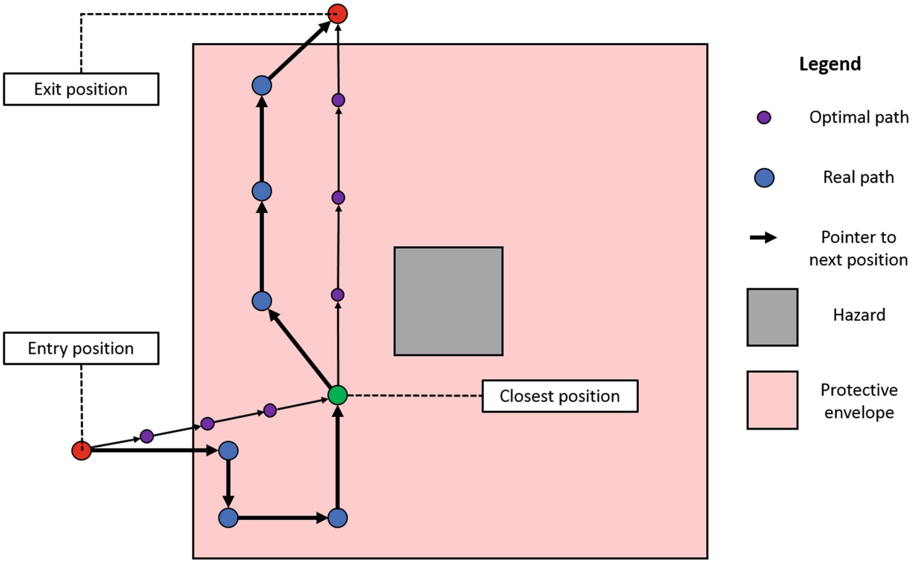


Fig. 6. Real path and direct path

6.2 Radar Plot

For each close call a radar plot is computed showing the weight values for velocity, duration, deviation, distance, and orientation. These weights, as explained next, visualize the severity of the different aspects that contributed to the close call event. The higher the value points in the radar plot, the more contributed the aspect to the endangerment of the resource. Velocity and length during the close call event (see Fig. 7) give a user a brief overview of a resource's safety performance. As suggested by [25] personalized feedback or other change (i.e., selection of other equipment or type, modification to site layout plan) can be issued and future performance monitored until the issue is resolved.

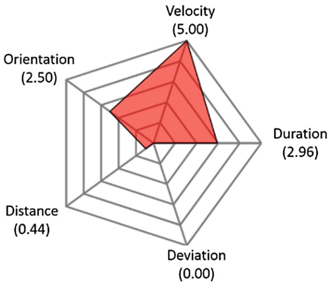


Fig. 7. Radar plot indicating factors leading to a close calls

6.3 Hazard Weights

The following introduces the formulas to calculate the weights (velocity, duration, deviation, distance, and orientation) (Fig. 8). While the original values for the weights can be based on historical data records, they can be adjusted over time. $Weight_{\max}$ refers to a maximum weight.

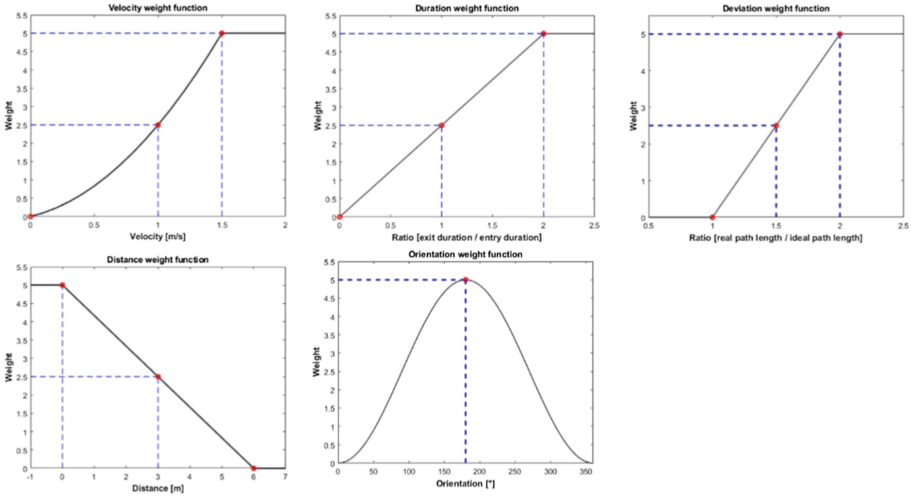


Fig. 8. Weight functions

6.3.1 Velocity Weight

The velocity weight for a close call is calculated by using the velocity weight function (Fig. 8), with the average velocity of the close call as an input. In addition to $Weight_{\max}$ the course of this function depends on the parameter $Vel_{\max}$ which represents the maximum velocity a vehicle or pedestrian is allowed to have. [36] points out that there is no common definition for safe velocities to operate construction equipment. The speed limits on construction sites depend on numerous factors like type of equipment or underground conditions [37]. In the following sections, $Vel_{\max}$ is assumed to be 1 m/s (or 3.6 km/h).

It is assumed that a velocity of 0 is always the safest and therefore the weight is set to 0 for all parameters of $Weight_{\max}$ and $Vel_{\max}$. Furthermore, moving with a velocity equal to the speed limit $Vel_{\max}$ is weighted with $\frac{Weight_{\max}}{2}$. Since the risk of severe injuries increases exponential the higher the velocity is, the weight function also increases exponentially. Moving with a speed of 150% of the allowable speed limit is rated with $Weight_{\max}$ and all velocities over this threshold alike. In brief, these conditions lead to three specific points:

- $P_0 = (0, 0)$
- $P_1 = \left(Vel_{\max}, \frac{Weight_{\max}}{2}\right)$
- $P_2 = (1.5 * Vel_{\max}, Weight_{\max})$

on the velocity weight function which is of the form $f(x) = ax^2 + bx + c$. Inserting these points into this function creates a linear system of equations which can be written as a matrix equation

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & Vel_{\max} & Vel_{\max}^2 \\ 1 & (1.5 * Vel_{\max}) & (1.5 * Vel_{\max})^2 \end{pmatrix} * \begin{pmatrix} c \\ b \\ a \end{pmatrix} = \begin{pmatrix} 0 \\ 0.5 * Weight_{\max} \\ Weight_{\max} \end{pmatrix} \tag{5}$$

and solved using the MATLAB® matrix division operation.

6.3.2 Duration Weight

The duration weight could be determined by using the duration of the close call alone. However, this might lead to a correlation between the size of a hazard envelope and the duration weight, as the risk of being longer inside a hazard increases with its size. Given that one of this research's aim is to quantify aspects that help to analyze pedestrian workers' behavior, it is more sensible to examine the ratio between entry duration and exit duration. This value could indicate if the worker noticed the hazard or if the worker took action accordingly and left the hazard area soon after sensing the danger. Combined with other values, for example the exit velocity, one can draw more conclusions about the incident.

The weight function (Fig. 8) is composed of a linear function for ratios from 0 to $Ratio_{\max}$ and a constant function with a value of $Weight_{\max}$ for all ratios above $Ratio_{\max}$. In the event of the entry duration being equal to the exit duration, the weight function returns half of $Weight_{\max}$.

Figure 8 displays the duration weight function for $Ratio_{\max} = 2$ so that it returns $Weight_{\max}$ once the entry duration is at least half as long as the exit duration. Let the ratio for the duration weight function be defined as

$$Ratio_{duration} = \frac{exitDuration}{entryDuration} \tag{6}$$

Then the weight function for duration is defined as

$$Weight_{Duration}(Ratio_{duration}) = \begin{cases} \frac{Weight_{\max * Ratio_{duration}}}{Ratio_{\max}}, 0 \leq Ratio_{duration} \leq Ratio_{\max} \\ Weight_{\max}, Ratio_{duration} \geq Ratio_{\max} \end{cases} \quad (7)$$

6.3.3 Deviation Weight

The ratio between the length of real path and optimal path (Fig. 8) is described as the deviation of a close call event. Since the ideal path of a close call event leads directly through three positions of the real path (namely: entry, closest, and exist points), the real path length is always greater than or equal than the ideal path length. Therefore, the ratio between these values is 0, if both lengths are equal. In this case the worker walked the ideal path and the deviation weight is set to 0. A weight of $W_{\max}$ is assigned if the actual walked path is twice as long as the ideal path length. Let $Path_{real}$ be the real path length and $Path_{opt}$ be the ideal path length. Then the deviation weight function can be defined as

$$Weight_{Dev}(Path_{real}, Path_{opt}) = \begin{cases} \left(\frac{Path_{real}}{Path_{opt}}\right) - 1; 1 \leq \left(\frac{Path_{real}}{Path_{opt}}\right) \leq 2 \\ Weight_{\max}; Path_{real} \geq 2 * Path_{opt} \\ undefined; Path_{real} < Path_{opt} \end{cases} \quad (8)$$

6.3.4 Distance Weight

For the computation of the distance weight of a close call event, the resources individual safe distances as well as the closest distance are required. There are three major cases to distinguish for the distance between two resources (see Fig. 9):

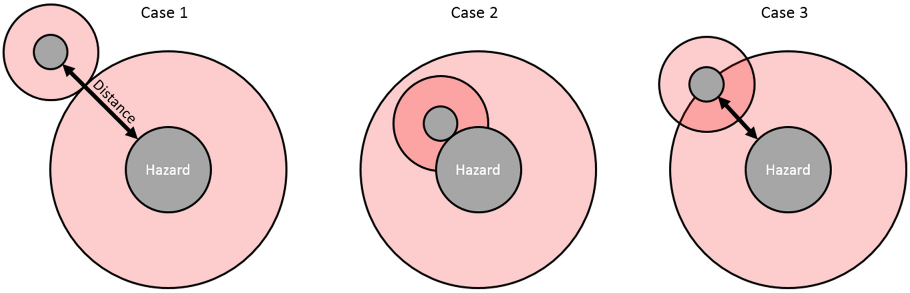


Fig. 9. Three cases for distance between two resources

- Case 1: The distance is equal to the sum of both safe distances. This is assumed to be the best case and a weight of 0 is assigned.
- Case 2: The distance is equal or smaller than 0 which is the case if the resource models overlap. This would be the worst case and is evaluated with $W_{\max}$.

- Case 3: The distance lies between the two cases mentioned above. In this case the assigned weight is between 0 and $W_{\max}$.

Let D_A and D_B be the assigned safe distances for resource A and resource B in meters with $D_A \leq D_B$, let d be the input distance and D_{sum} be the sum of D_A and D_B. The distance weight function is partially defined as a linear function for distances between 0 and D_{sum}, composed with a constant function of $Weight_{\max}$ for all distances that are smaller than 0 and another constant function of 0 for all distances greater than D_{sum}. The slope of the linear function is equal to $-\frac{Weight_{\max}}{D_{sum}}$. In summary, the weight function can be written as

$$W_{Distance}(d, D_{sum}) = \begin{cases} Weight_{\max}; d \leq 0 \\ -\frac{Weight_{\max}}{D_{sum}} * d + Weight_{\max}; 0 < d < D_{sum} \\ 0; d \geq D_{sum} \end{cases} \tag{9}$$

Figure 8 displays the distance weight function for safe distances of $D_A = 1m$ and $D_B = 5m$ whereas the value for d ranges from −1 to 7.

6.3.5 Orientation Weight

Computed orientations, as shown earlier, range from 0 to 360°. Using the average orientation over all buffer events is not feasible as potential left-side and right-side orientations would cancel each other out (average of 90° and 270° is 180°). Therefore, the orientation weight W_{orient} depends on three values:

- O_{entry}: Orientation at entry event buffer.
- O_{exit}: Orientation at exit event buffer.
- $O_{closest}$: Orientation at closest position event buffer.

Separate orientation weight values for each of these three values are calculated. Evaluating the orientation is then a matter of perspective. Weighting hazards appearing from the front (around 0°) can help to find inattentive workers, while hazard behind a worker can pose a dangerous threat even to very cautious workers. Therefore, unless other methods are used to track whether a human has recognized a hazard or not, the evaluation of orientation depends on the users' personal preferences. In the presented scenario, hazards from behind will be evaluated as more dangerous.

Then the function is based on the sinus function, which is translated upwards on the y-axis by 1, then translated vertically to the left on the x-axis by $\frac{3\pi}{2}$, then stretched horizontally by a factor of $\frac{180}{\pi}$ and then stretched vertically by a factor of $\frac{Weight_{\max}}{2}$. To make sure that the absolute orientation weight is not greater than $Weight_{\max}$ the weights for closest orientation, entry orientation and exit orientation are all divided by three and then added together. Let the function for single orientation weights be defined as:

$$Weight(orient) = \left(\sin\left(\frac{orient * \pi}{180} + \frac{3\pi}{4}\right) + 1\right) * \frac{Weight_{\max}}{2} \tag{10}$$

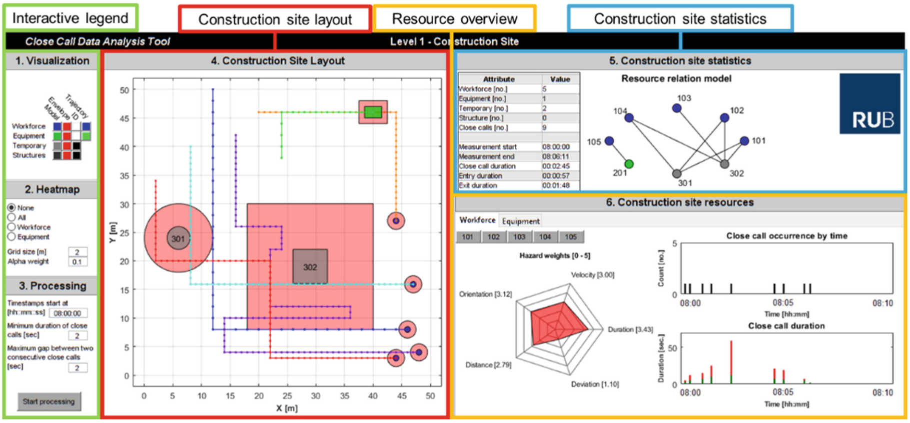
$$Weight_{Orient} = \frac{Weight(Orient_{closest}) + Weight\left(Orient_{entry}\right) + Weight(Orient_{exit})}{3} \quad (11)$$

There is also the possibility to rate both, hazards from behind and from the front with high weights. However, this would cause the values to lose their informative value since the weight would be the same for incautious workers which do not recognize a hazard as well as for workers which could not see the hazard from behind. Im brief, let the users configure the tool based on their personal preferences.

6.4 Visualization

The computational analysis of the gathered and fused data starts with the examination of all single event buffers. From there, it abstracts and combines these information into more general statistics (see Fig. 3). The Guided User Interface (GUI) displays the construction site layout on a map, general construction site statistics, and an overview to all construction resources, separated by type. A heatmap, if selected by the user, shows the location of close calls.

The GUI is shown in Fig. 10 contains the general close call performance information for a construction site. It covers statistical data as well as a brief overview on all resources being present at the construction site and involved in close calls. The GUI might be used by management to derive a quick performance overview on close calls at one construction site.



◉ None
○ All
○ Workforce
○ Equipment

Grid size [m] 2
Alpha weight 0.1

Timestamps start at [hh::mm::ss] 08:00:00
Minimum duration of close calls [sec] 2
Maximum gap between two consecutive close calls [sec] 2

Attribute	Value
Workforce [no.]	5
Equipment [no.]	1
Temporary [no.]	2
Structure [no.]	0
Close calls [no.]	9
Measurement start	08:00:00
Measurement end	08:06:11
Close call duration	00:02:45
Entry duration	00:00:57
Exit duration	00:01:48

Fig. 10. GUI level 1 – construction site (Color figure online)

Since this GUI window is the first that opens in the current close call analysis, a user may configure the processing parameters (see an interactive legend in Fig. 10). Results that are illustrated in the other subparts of the figure (titled with 4.–6.) change accordingly. In subpart 1, an interactive legend enables the user to hide and show the boundaries of objects that are present on site. The user can enable or disable the visualization of the corresponding protective envelopes, identification numbers (IDs), and trajectories (colorized by resource) on a construction site layout map. The latter is built in advance from a regularly updated site layout plan using BIM [23, 38].

In subpart 2, the interactive legend panel allows changes to the close call heatmap's grid size. The construction site layout window (see subpart 4) displays the heatmap for the resources workforce and equipment separately. Inside the processing configuration panel (subpart 3) are three editable fields to influence the computational data analysis (e.g., the timestamp to begin the analysis). Then a user selects the minimum duration of a close call event. It defines how long a close call has to be included in the analysis. Close calls that only last for one second might frequently be found in the results, but may not provide valuable information. In fact, in situations when equipment frequently passes by pedestrian workers, many of these might be interpreted erroneous or irrelevant. On the other hand, allowing a user to define a gap value (the maximum time between two consecutive close call events) adds or limits granularity in the close call analysis.

The GUI further consists of a tab for construction site statistics (see subpart 5). It gives an overview to the analysis to all close call events that happened at this construction site. The resource relation model visualizes those resources that are most often involved in close calls. The example shown, shows close calls that involve workforce (1xx) with equipment (2xx) or hazardous objects (3xx). No close call between equipment and hazardous objects was observed in the artificially generated data set.

A radar plot in subpart 6 of the figure shows the calculated weights for velocity, duration, deviation, distance, and orientation for all workers. In this example, 4 out of 5 weights pose a flag and may require safety management to act upon. Additionally, two plots for occurrences of close calls by time and duration of close call events give a brief overview of critical periods of the observed time interval. The red bars on the duration plot represent the exit duration and the green bars show the entry duration of close calls.

7 Verification of Method

A simulated scenario verified the close call analysis algorithm. An artificial data was manually generated. The data set included five workers that traverse a construction site in a continuous manner, facing two temporary static hazards and one dynamic vehicle. Each worker simulates a behavior which should raise one of the different hazard weights. To raise the orientation weight value for a worker, for example, the vehicle creates a close call in a workers' blind space. All trajectories are straight lines. This permits simplicity in the verifying process of the algorithm. A heatmap displayed in the GUI further allows a user to identify the close calls.

In the simulated scenario, one pedestrian worker (A) traversed the site using a speed of 2 m/s (at a maximum allowable speed limit of 1 m/s). A second pedestrian worker (B) kept a too short distance towards the hazards (301 and 302). A third pedestrian worker (C) simulated a behavior which should result in a high deviation weight. The duration weight was tested by pedestrian worker (D). Pedestrian worker (E) was confronted with a traversing vehicle (F) to verify the orientation weight function. The heatmap functionality was verified by comparing the trajectories with the hazard locations on the map (see Fig. 11).

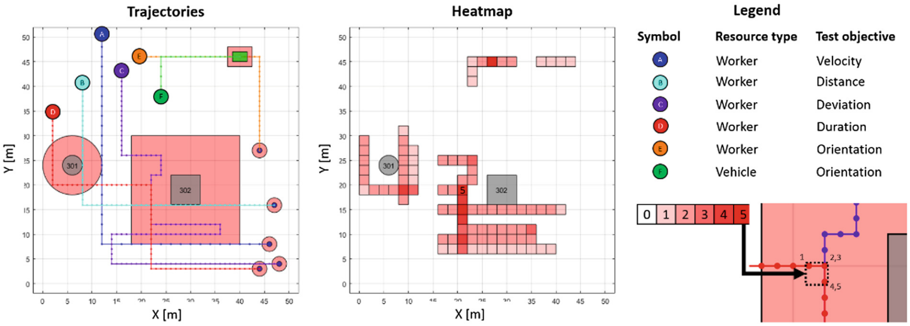
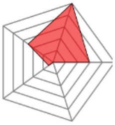
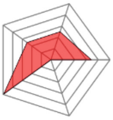
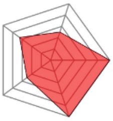
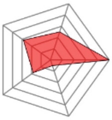
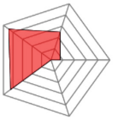
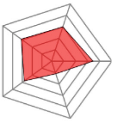


Fig. 11. Verification of close call analysis algorithm using an artificial construction scenario (Color figure online)

The weight radar plots for all resources are displayed in Table 1. Resources A, B and D showed expected results and verify the functionality of the velocity, distance, and duration weights. In contrast to the other resources, C shows two raised weights for deviation and duration. As the deviation value quantifies the straying of the worker from an ideal path and the duration value increases with a longer exit duration, a raised deviation weight might tend to be accompanied by a raised duration weight. In contrary, the radar plot of resource D shows a sole raised duration weight. Therefore, a mutual correlation between these values can be excluded. Resource E shows two raised weight values as well. This can be explained as: (a) vehicle and pedestrian worker do not have a large safety distance and (b) the vehicle stopped right behind the worker with a distance of close to zero meters. In theory, each one of the recorded close call events should be followed up. However, a user in a realistic scenario may need to set preferences on the more severe close calls. According to the initial findings in a simulated test environment, weight values of approximately 4 or higher, require such detailed follow-ups.

Table 1. Weight radar plots and values for every resource and average team performance

Category			**Resource**			**Team**
	A	B	C	D	E	
Radar Plot						
Velocity	**5,00**	2,50	2,50	2,50	2,50	3,00
Duration	2,96	3,42	**5,00**	**5,00**	0,78	3,43
Deviation	0,00	0,00	**5,00**	0,51	0,00	1,10
Distance	0,44	**4,93**	2,50	1,84	**4,26**	2,79
Orientation	2,50	2,50	3,26	2,86	**4,46**	3,12

8 Results to Validation

To validate the close call data analysis algorithm, two real datasets from two different construction sites were analyzed. The following sections cover the pedestrian workers individual performances and the overall construction site safety performance (incl. the team). Discussions on proposals for improvements on the construction site follow.

8.1 Experiment 1: Building Construction Site

The first dataset was gathered from a building construction site where several pedestrian workers were present at an elevated work level. A restricted workspace was located inside it. Although the protective guardrails around the leading edges met the required safety standards, the present supervisor estimated it as insufficient (asking his and subcontracted personnel "to stay away from the edges"). One of the particular concerns was the arrival of a subcontractor. A new work crew for tying rebar yet had to familiarize itself with the work environment (including work at height). Therefore, the close call analysis algorithm aimed at analyzing the trajectories of three of the subcontracted workers for potential close calls and unauthorized entry into the restricted work space.

As shown in Fig. 12 (see the grey objects) the restricted space and the leading edges were modelled as individual objects using BIM. UWB served as sensing technology for recording the trajectories of the personnel. The information in Fig. 12 displays also individual trajectories and resulting heatmaps for every worker. The images indicate several close calls, mostly towards the southern and eastern part of the construction site. Interestingly to note are the green tiles, also visualized in Fig. 12. They indicate that a person entered the material storage area. Since it was the material manager (tag ID: 0000080E), there was no violation. Worker 000065BB once passed by the restricted work space. As shown, the use of sensing technology, data analysis and visualization offers also the option of positive feedback. Since most workers fear strict retaliation, programs can be developed that heighten workers' morale. As a consequence, the responsible safety personnel on site could be advised to inspect the leading edges that are marked in red in Fig. 12. Showing an illustration like Fig. 12 (object locations with close calls are highlighted in red) could even be sown to the workforce in Job Hazard Analysis (JHA) or toolbox talks ahead of task execution. While providing active feedback with realistic data from the same construction site has the potential to strengthen workers' risk awareness quickly, yet future research has to validate this assumption.

The analysis of several of the generated hazard weight radar plots for the pedestrian workers give further insights into the observed close calls. Table 2 displays the individual workers' hazard weight radar plots and the team's performance. The worker with the ID: 00000BC6 shows higher hazard weights than most other workers. Although the data indicates only two close calls nearby, they must rather have been serious close calls (high speed and very close to leading edges). In such a case, the worker could be instructed first, then pulled temporarily from work and provided with proper instructions.

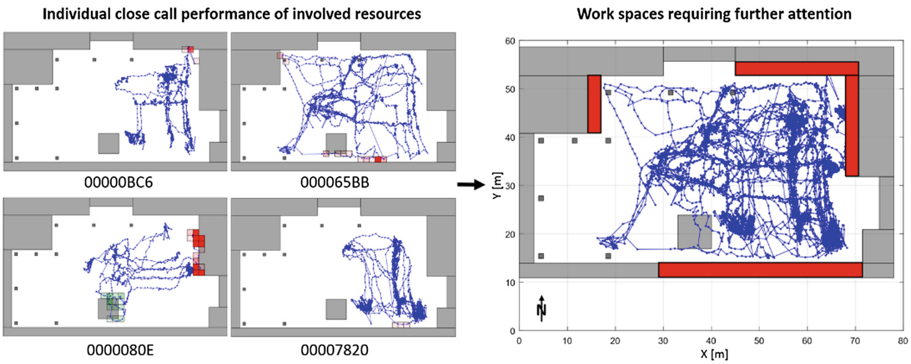
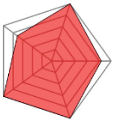
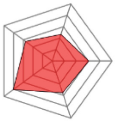
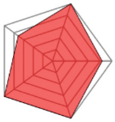
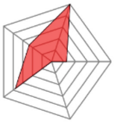
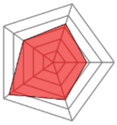

Individual close call performance of involved resources
00000BC6
000065BB
0000080E
00007820
Work spaces requiring further attention
Y [m]
X [m]
0
10
20
30
40
50
60
70
80
N

Fig. 12. Heatmaps identify close calls in a BIM-based site layout (extracted for each resource and construction sites from the GUI in the order of appearance from left to right) (Color figure online)

Table 2. Weight radar plots and values for every resource and average team performance

Category	Resource				Team
	00000BC6	000065BB	0000080E	00007820	
Radar Plot					
Velocity	**5,00**	2,33	0,85	**5,00**	3,29
Duration	**4,17**	3,05	2,47	1,25	2,73
Deviation	**5,00**	2,77	**5,00**	0,00	3,19
Distance	**5,00**	**4,04**	**5,00**	3,78	**4,46**
Orientation	3,56	2,78	3,33	2,34	3,00

8.2 Experiment 2: Infrastructure Construction Site

A second realistic trial of the close call analysis algorithm utilized data from a large infrastructure construction site. In the experimental testbed 4 pedestrian workers, 1 tractor, and 1 mobile crane operated conjointly in a confined space (an excavated pit). While the original data analysis was performed by Cheng et al. (2011), the goal of this test was to find close calls between the pedestrian workers and the moving construction equipment or parts of it (the body of a tractor and the moving load of mobile crane). The potential hazard of a pedestrian worker being pinned by the rotating body of the mobile crane was not analyzed, because its outriggers were safely guarded. Similar to the first experiment, the results show the individual trajectories, heatmaps (see Fig. 13), and hazard weight radar plots (see Fig. 14).

Both pedestrian workers with ID: 00000BC6 and ID: 0000080E were not involved in a close call. Worker with ID: 00005AA1 came several times very close to or under

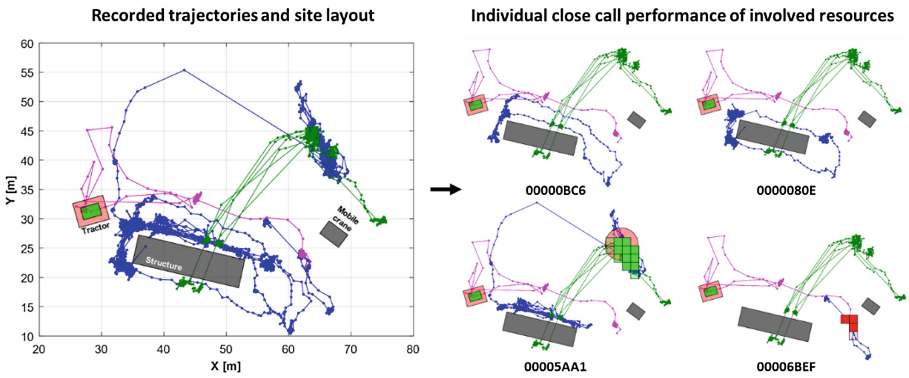


Fig. 13. Construction site layout and trajectories (equipment movement in green and pedestrian workers in blue color) (Color figure online)

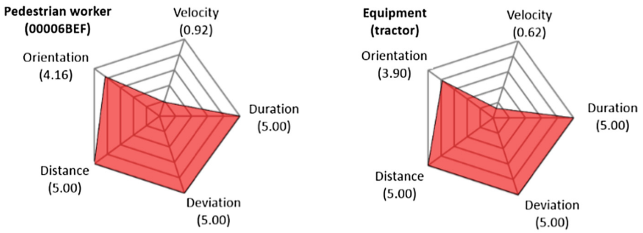


Fig. 14. Hazard weight radar plots

the swings the mobile crane performed. However, as [39, 40] have shown on the same data set, the worker was authorized to work near the operating crane. Therefore, the tiles are marked green to allow the worker to detach or attach loads to the crane hock.

The trajectory of the pedestrian worker with ID: 00006BEF, however, collided with the path of the tractor that delivered material into the pit. The tractor's and the pedestrian worker's hazard weight radar plots (See Fig. 14) show nearly matching values for 5 of the observed values. Although these close calls were revealed, they were not severe as both resources moved with very low velocities (≤ 1 m/s). One could argue that the pedestrian worker operated as a temporary flagman, guiding the vehicle into a confined space inside the excavated pit.

9 Discussion, Limitations, and Conclusion

This study presented an algorithm for the quantitative analysis of close call events in construction. A process of collecting trajectory data to valuable construction resources was introduced and a graphical user interface was presented that provides safety personnel with automatically generated safety information on close calls. The algorithm was then successfully tested in simulated and realistic work environments.

Although the develop algorithm provides useful information on both artificial and real trajectories, the performed calculations are based on several assumptions. They rely in particular on commercially-available real-time location sensing technology. Known measurement errors may not qualify these for application in the harsh construction environment. Though [31] demonstrated that errors with UWB, for example, can be below 1 m for each positional data log, complex sensor infrastructure to set up and maintain (i.e. ensuring signal strength while filtering multipath) might be unfavorable looked upon when considering economical application in construction. Any technology must also withstand ethical concerns of tracking workforce and be effective in acquisition, use, and maintenance. While the latter issue could be solved by targeting worthwhile logistics applications at the same time, indoor work environments – where most location tracking technologies will not work – demand new sophisticated solutions.

On a similar note, the developed algorithm considers trajectory-related information only. Though this adds new functionality to existing close call management processes, additional research is necessary. For example, the presented hazard weight calculations are based on simplified assumptions. Field-based observations are likely necessary to complement the terms and calibrate the weights accordingly. Data fusion including new data points from proximity alert sensors [41–43], could serve future research agendas well. A port to safe test bed environments within virtual reality environments would enhance more realistic education and training scenarios, providing users with much needed personalized feedback.

References

1. Bird, F.E., Germain, G.L., Bird Jr., F.E.: Practical Loss Control Leadership. Revised edn. Intl. Loss Control Inst. (1996). ISBN-13: 978-0880610544
2. Cavalieri, S., Ghislandi, W.M.: A conceptual structure for the use of near-misses properties. IFAC **39**(3), 81–86 (2006)
3. NIOSH homepage. https://www.cdc.gov/niosh/face/inhouse.html. Accessed 20 Dec 2017
4. Teizer, J.: Right-time vs. real-time pro-active construction safety and health system architecture. Constr. Innov. Inf. Process Manag. **16**(3), 253–280 (2016)
5. Marks, E., Teizer, J.: Method for testing proximity detection and alert technology for safe construction equipment operation, construction management and economics. Occup. Health Saf. Constr. Ind. **31**(6), 636–646 (2013). http://www.tandfonline.com/doi/abs/10.1080/01446193.2013.783705. Taylor & Francis, Special Issue
6. CII: Near Miss Reporting to Enhance Safety Performance, The University of Texas at Austin (2014)
7. Cambraia, F.B., Saurin, T.A., Formoso, C.T.: Identification, analysis and dissemination on near misses: a case study in the construction industry. Saf. Sci. **48**, 91–99 (2010)
8. Heinrich, H.W.: Industrial Accident Prevention: A Scientific Approach. McGraw-Hill, New York (1931)
9. Smallwood, J., Emunze, F.: Towards zero fatalities, injuries and disease in construction. Procedia Eng. **164**, 453–460 (2016)
10. Gnoni, M.G., Lettera, G.: Near-miss management systems: a methodological comparison. J. Loss Prev. Process Ind. **25**, 609–616 (2012)

11. Gnoni, M.G., Saleh, J.H.: Near-miss management systems and observability-in-depth: handling safety incidents and accident precursors in light of safety principles. Saf. Sci. **91**, 154–167 (2017)
12. Raviv, G., Fishbain, B., Shapira, A.: Analyzing risk factors in crane-related near-miss and accident reports. Saf. Sci. **91**, 192–205 (2017)
13. Teizer, J., Cheng, T.: Proximity hazard indicator for workers-on-foot near miss interactions with construction equipment and geo-referenced hazard areas. Autom. Constr. **60**, 58–73 (2015)
14. Lu, M., Cheung, C.M., Li, H., Hsu, S.: Understanding the relationship between safety investment and safety performance of construction projects through agent-based modeling. Accid. Anal. Prev. **94**, 8–17 (2016)
15. Li, H., Lu, M., Hsu, S.-C., Gray, M., Huang, T.: Proactive behavior-based safety management for construction safety improvement. Saf. Sci. **75**, 107–117 (2015)
16. Hilfert, T., Teizer, J., König, M.: First person virtual reality for evaluation and learning of construction site safety. In: 33rd International Symposium on Automation and Robotics in Construction, Auburn, Alabama, USA (2016)
17. Wu, W., Yang, H., Chew, D.A.S., Yang, S., Gibb, A.G.F., Li, Q.: Towards an autonomous real-time tracking system of near-miss accidents on construction sites. Autom. Constr. **19**, 134–141 (2010)
18. Teizer, J., Cheng, T., Fang, Y.: Location tracking and data visualization technology to advance construction ironworkers' education and training in safety and productivity. Autom. Constr. **35**, 53–68 (2013)
19. Rozenfeld, O., Sacks, R., Rosenfeld, Y., Baum, H.: Construction job safety analysis. Saf. Sci. **48**, 491–498 (2010)
20. Zhang, S., Teizer, J., Lee, J.-K., Eastman, C., Venugopal, M.: Building information modeling (BIM) and safety: automatic safety checking of construction models and schedules. Autom. Constr. **29**, 183–195 (2013)
21. Zhang, S., Boukamp, F., Teizer, J.: Ontology-based semantic modeling of construction safety knowledge: towards automated safety planning for job hazard analysis (JHA). Autom. Constr. **52**, 29–41 (2015)
22. Lu, Y., Li, Q., Zhou, Z., Deng, Y.: Ontology-based knowledge modeling for automated construction safety checking. Saf. Sci. **79**, 11–18 (2015)
23. Schwabe, K., König, M., Teizer, J.: BIM applications of rule-based checking in construction site layout planning tasks. 33rd International Symposium on Automation and Robotics in Construction, Auburn, Alabama, USA (2016)
24. Zhang, S., Teizer, J., Pradhananga, N., Eastman, C.M.: Workforce location tracking to model, visualize and analyze workspace requirements in building information models for construction safety planning. Autom. Constr. **60**, 74–86 (2015)
25. Golovina, O., Teizer, J., Pradhananga, N.: Heat map generation for predictive safety planning: preventing struck-by and near miss interactions between workers-on-foot and construction equipment. Autom. Constr. **71**, 99–115 (2016)
26. Pradhananga, N., Teizer, J.: Automatic spatiotemporal analysis of construction site equipment operations using GPS data. Autom. Constr. **29**, 107–122 (2013)
27. Vasconcelos, B., Barkokébas Jr., B.: The causes of work place accidents and their relation to construction equipment design. Procedia Manuf. **3**, 4392–4399 (2015)
28. Hinze, J.W., Teizer, J.: Visibility-related fatalities related to construction equipment. Saf. Sci. **49**, 709–718 (2011)

29. Ranney, J.M., Zuschlag, M.K., Morell, J., Coplen, M.K., Multer, J., Raslear, T.G.: Evaluations of demonstration pilots produce change: fourteen years of safety-culture improvement efforts by the federal railroad administration. TR News – Railroads Res. Shar. Track **286**, 28–36 (2013)
30. Hallowell, M.R., Hinze, J.W., Baud, K.C., Wehle, A.: Proactive construction safety control: measuring, monitoring, and responding to safety leading indicators. J. Constr. Eng. Manag. **139**, 04013010 (2013)
31. Cheng, T., Venugopal, M., Teizer, J., Vela, P.A.: Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments. Autom. Constr. **20**(8), 1173–1184 (2011)
32. Siebert, S., Teizer, J.: Mobile 3D mapping for surveying earthwork projects using an unmanned aerial vehicle (UAV) system. Autom. Constr. **41**, 1–14 (2014). https://doi.org/10.1016/j.autcon.2014.01.004
33. Kim, C., Haas, C.T., Liapi, K.A., McLaughlin, J., Teizer, J., Bosche, F.: Rapid human-assisted, obstacle avoidance system using sparse range point. In: Proceedings of the 9th Biennial ASCE Aerospace Division International Conference on Engineering, Construction, and Operations in Challenging Environments, League City, Houston, Texas, pp. 115–122 (2004)
34. Teizer, J.: Safety 360: surround-view sensing to comply with changes to the ISO 5006 earth-moving machinery - operator's field of view - test method and performance criteria. In: Proceedings of the 32nd International Symposium on Automation and Robotics in Construction, Oulu, Finland (2015)
35. BG Bau homepage. http://www.bgbau-medien.de/struktur/inh_baus.htm. Accessed 30 Nov 2017
36. OSHA homepage: Evaluation of what is considered a safe speed to operate a powered industrial truck, Occupational Safety and Health Administration. https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=INTERPRETATIONS&p_id=24995. Accessed 30 Nov 2017
37. Kamat, V.R., Martinez, J.C.: Visualizing simulated construction operations in 3D. J. Comput. Civ. Eng. **15**(4), 329–337 (2001)
38. Krepp, S., Jahr, K., Bigontina, S., Bügler, M., Borrmann, A.: BIMsite - towards a BIM-based generation and evaluation of realization variants comprising construction methods, site layouts and schedules. In: Proceedings of the EG-ICE Workshop on Intelligent Computing in Engineering, Krakow, Poland (2016)
39. Yang, J., Vela, P.A., Teizer, J., Shi, Z.K.: Vision-based crane tracking for understanding construction activity. ASCE J. Comput. Civ. Eng. **28**(1), 103–112 (2014)
40. Cheng, T., Teizer, J.: Modeling tower crane operator visibility to minimize the risk of limited situational awareness. ASCE J. Comput. Civ. Eng. **28**(3), 04014004 (2014). http://ascelibrary.org/doi/abs/10.1061/(ASCE)CP.1943-5487.0000282
41. Teizer, J., Allread, B.S., Fullerton, C.E., Hinze, J.: Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system. Autom. Constr. **19**(5), 630–640 (2010). https://doi.org/10.1016/j.autcon.2010.02.009
42. Luo, X., Li, H., Huang, T., Rose, T.: A field experiment of workers' responses to proximity warnings of static safety hazards on construction sites. Saf. Sci. **84**, 216–224 (2016)
43. Yang, K., Ahn, C.R., Vuran, M.C., Aria, S.S.: Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit. Autom. Constr. **68**, 194–202 (2016)

Towards a Data-Driven Approach to Injury Prevention in Construction

Junqi Zhao and Esther Obonyo

The Pennsylvania State University, State College, PA 16802, USA
jzz30@psu.edu, eao4@engr.psu.edu

Abstract. The research discussed in this paper is part of a project directed at increasing productivity in construction through mitigating the risk of Musculoskeletal Disorders (MSD). Postures and activities recognition through motion capturing techniques have shown promising potential for monitoring, assessing, and reducing such risks. Current motion sensing systems require a complex whole-body senor placement to capture and recognize construction activities, which limits the practicality and requires great computational effort. This challenge can be addressed through using a machine learning approach that recognizes specific activities from human motion data. The feasibility of reducing the computational effort through using fewer sensors rather than whole-body sensor placement was assessed through a case study. Five sensors were placed in targeted motion areas. The authors propose a novel automatic model configuration process to improve recognition performance under the selected sensor placement. It is based on designing optimal combination of data segmentation window size, feature sets, and classification algorithms for a specific set of injury-prone construction activities. The proposed approach achieved an average overall recognition accuracy of 0.81 and 0.74 for two sets of activities. The recognition model operation time is also reduced to less than 0.01 s under the proposed approach. In this initial case study, the model configuration process was developed iteratively based on the output from the test case. In subsequent efforts, the authors will develop a generic activity recognition model with predefined rules and criteria. This will further accelerate and automate the model configuration process.

Keywords: Sensors · Machine learning · Computational effort
Activity recognition · Injury prevention

1 Introduction

Construction operations are usually complex and physically demanding activities [1], which tends to expose construction workers to awkward working postures. The prolonged awkward working postures and activities might cause not only the instantaneous injury like fall [2] but, more frequently, the accumulative non-fatal injuries, particularly the Musculoskeletal Disorders (MSD) [3]. Such disorders are becoming increasingly prevalent and induce both productivity and economic loss. Between 2011 and 2014, the MSD-related injuries in construction increased by 14.6% (Data Tools of Bureau of Labor Statistics 2016). The reported MSD rate in construction was more than double the average

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 385–411, 2018.
https://doi.org/10.1007/978-3-319-91635-4_20

rate for all occupations. MSD also accounts for over 37% of all injuries to construction trade workers that result in days away from work [3]. The problem could be worse given the fact that some work-related injuries go unreported [4].

Effective MSD prevention needs a clear understanding of its development. Multiple risk factors have been identified, ranging from external (awkward postures, forceful exertions), internal (joint moments and muscle forces), to metabolic demands (muscle fatigues) [5–7]. In construction, the awkward working postures and activities have been repeatedly reported as a prominent risk factor for workers [2, 8, 9], which involves body positions deviating significantly from the neutral position and potentially lead to cumulative trauma disorders if maintained overtime. Exemplary activities include overhead working, kneeling, back bending forward, squatting, neck bending, and reaching [2, 10]. Capturing workers' posture and activity information will have significant implication for construction safety management and productivity improvement. The captured workers' working posture type, duration, and activity information would be helpful for proper training, redesigning the workspace, and triggering real-time intervention to reduce injury risks [2]. The collected activity information can also benefit the individual worker by improving their safety awareness and realize self-management of awkward activities [11]. Besides, the properly scheduled break and recovery time can also help to mitigate injury risks and improve productivity [12]. Therefore, it is meaningful to accurately monitor the occurrence of the targeted activities in construction.

Various ergonomics assessment tools and motion capture systems have been adopted for monitoring and assessing the construction activities. Particularly, the emerging motion sensing techniques have dramatically improved the activity and posture data capturing efficiency and accuracy compared with traditional manual observation approach [13–15]. Using bridge design as an analogy, the kind optimal configurations for the use of low cost sensors to requires a holistic assessment of factors such as soil and condition, bridge span type (cable-stayed or suspension) and bridge length seismic zone. These are not standalone factors and as such, the configuration of sensors on bridge should take into account their synergistic contribution to the optimal performance of the bridge. To address the need to holistically assessing the deployment context, some solutions capture subjects' motion data, through full body sensor placement approach that are either optical or vision-based [15]. In order to accurately quantify subjects' activities, captured motion data are transformed and reconstructed into a full-body model to provide meaningful physical information, such as full body joint movement angles [11], muscle activities [21], or gait status [16]. Combinations of complex whole-body motion sensing systems and domain knowledge have demonstrated great potential for physical condition monitoring and achieved exceptional high accuracy for posture measurement. However, the current approaches are usually applied for a single subject in the controlled lab setting, the motion data are processed "off-line" after data collection, which makes it not applicable for scalable application in construction. This can be attributed to the: (i) computational cost from the great volume of whole-body sensing data and the subsequent complex data processing process [2]; (ii) vulnerability of vision and optical based sensors to complex outdoor lighting condition [15]; (iii) strict calibration requirement for accurate posture measurement of motion sensors [17, 18]; (iv) potential intrusiveness for workers' activity due to whole-body sensor placement [15, 19]; and (v) high cost of sensors in scalable application [19, 20]. There is a need for an activity recognition approach that can be easily deployable in real jobsite conditions through taking into account both the technical complexities (such as sensor requirements and computational cost) and practical usability (such as non-intrusiveness). There are two significant challenges associated with the deployment of whole-body sensor systems: (1) they typically rely on the use of complex and often more expensive sensors, and; (2) the data derived are more difficult to work with respect to managing the sheer volume and computational intensity required to process the data.

Some of the challenges related to the deployment of whole body systems can be addressed through using a limited number of low-cost sensors that feed data into a machine-learning enhanced data synthesis and processing backend system. In the context of detecting movements inherent in routine human activities, the successful design and deployment of a machine learning-based activity recognition approach predicates on the optimal configuration of a several of interacting factors. For the application context discussed in this paper, these factors include sensor placement, data segmentation window size, classification algorithms, and feature sets. The goal here is achieving the optimal trade-off between the need to increase activity recognition accuracy and minimizing computational efforts/costs. Although there have been some examples in which this need was addressed through the use of pre-defined parameters [2, 8, 21, 22], it limits the applicability of the resulting activity recognition model to a diverse set of users performing even when they are performing similar activities.

Because the optimal design of window size and recognition algorithms varies between activities, a predetermined configuration may not achieve the best performance for different activity sets [23]. The authors proposes to address this challenge thought using a rapid and automatic process to configure the design factors. This approach transcends factors such as the window size as done by Banos, Galvez, Damas, Pomares and Rojas [1], and includes other interacting factors such as feature sets and classification algorithm. This approach would result in the use of fewer sensors, which in turn will minimize both the effort invested in developing a sensor placement plan and the computational efforts, required to for the model configuration process. The proposed approach will increase the effectiveness, usefulness, and scalability of deploying low cost sensors in injury prevention for construction workers. The findings can be transposed to other applications areas outside of the construction job site that are increasingly using motion detection for proactively monitor the risk of injuries such as sports and fall detection solutions for the elderly.

The efficiency and effectiveness of the proposed model were assessed through evaluating the window size for data segmentation, classification algorithms, and feature sets. The feasibility of reducing the computational effort by using fewer sensors and configuring a high-performance recognition model was assessed through a test case. The remainder of the paper is organized as follows. A review of related work is provided in Sect. 2, the proposed process and test case are specified in Sect. 3, the result data analysis and discussion are in Sect. 4, followed by the conclusion, limitation, and future work in Sect. 5.

2 Theoretical Background

The most widely used methods for capturing workers' activity data on construction site are visual observation and remote sensing [13]. Visual observation predicate on ergonomists' personal judgment. They tend to be time-consuming [15] and adversely influenced by subjective bias [14] and might lead to an inaccurate assessment even for trained ergonomists [13]. Sensing techniques offer an automatic and more accurate approach and have been used, for example, in behavior recognizing, positioning [1, 24], to muscle load estimation [10]. There are some studies directed at addressing the barriers to their widespread adoption that were highlighted in the previous section.

The potential of Inertia Measurement Units (IMU) sensors is exemplified in posture and movement tracking applications for optimal performance in athletics, as well as healthcare [25]. The miniature and wireless IMU sensors make them easily attachable to the human body or cloth without interfering with activities. The IMU sensors could provide high-frequency real-time data reflecting human body kinetic condition and be deployed in scale with reasonable cost. In the construction industry, activity monitoring with IMU has also been identified to be emerging trend [14, 26]. It has been used as a complementary measurement for vision-based sensing [27]; measuring gait stability [28]; and detecting/classifying activities [29, 30].

Data processing approaches for the sensed motion data can be categorized into Knowledge-Driven (KD) and Data-Driven (DD) approaches [31]. Under the Knowledge-Driven approach, activity and physical models are construed based on domain knowledge. In comparison, the DD approach mainly focuses on the sensing data patterns characterizing the targeted activities or physical status. Both approaches have been adopted for processing construction worker's activity and physical condition data collected from various sensors as shown in Table 1.

KD approach is widely used from the review. The repeatedly adopted domain knowledge includes the ergonomic standards from ISO, threshold algorithms for unsafe postures, and biomechanical information, which enables estimation of internal body load and muscle fatigue. Such approach is semantically clear and logically elegant [31], which interprets collected activity data into clinically meaningful information. However, such approach relies on high-quality measurement. One exemplary case is modeling the worker's awkward activities through transforming IMU sensor data into joint angles. This could be realized by using raw sensor output [10, 26], while it might suffer from data quality issues due to sensor drift and noises. More stable measurement result can be achieved by transforming raw data into Euler [2] or Quaternion angles [11] after applying sensor fusion techniques, such as Kalman filter [41], while such techniques require either computational complex data processing process or sophisticated sensors with on-board filters. Besides, the accuracy of the measurement depends on not only high-quality sensor signal but also strict implementation assumptions, such as a sensor-to-segment orientation and axis alignment for joint angle measurement, which warrants the controlled manual calibration or identifying and calibrating axes in a complex computational process as proposed in [18]. The KD approach has shown promising results for activity recognition, while it would be challenging in the scalable

Table 1. Activity monitoring with sensing techniques for construction workers

Data processing	Author	Data collection	Data analysis
Knowledge-Driven	Valero, Sivanathan, Bosché and Abdel-Wahab [10] Golabchi, Han, Fayek and AbouRizk [32] Han and Lee [15] Wang, Dai, Ning, Dong and Wu [24]	Vision sensors (RGB-D, cameras, or Vicon) to capture whole-body motion data	Reconstructing 3D whole-body movement models; assessing joint movement angles with ergonomic standards; predicting internal body forces; identifying predefined unsafe behaviors
	Yan, Li, Li and Zhang [11] Keyserling, Brouwer and Silverstein [33] O'brien [34] Fang, Cho, Druso and Seo [35] Valero, Sivanathan, Bosché and Abdel-Wahab [10] Kim, Ahn, Stentz and Jebelli [16] Alwasel, Sabet, Nahangi, Haas and Abdel-Rahman [36]	IMU-based wearable system on body parts or whole-body	Transforming captured data into body joint angles or indexes reflecting body condition, such as balance; comparing with ergonomic standards or threshold algorithms to identify risky activities
	Ravi, Dandekar, Mysore and Littman [1] Wang, Dai, Ning, Dong and Wu [24] Wang, Chen, Zhao, Dai, Zheng and Wu [37]	Ultra-Wideband, physiological sensors are used for location and physical condition	The captured data were transformed into physiological conditions, such heart activity, and muscle fatigue, which are compared with predefined standards and reflect the riskiness of potential injuries

(*continued*)

Table 1. (*continued*)

Data processing	Author	Data collection	Data analysis
	Antwi-Afari, Li, Edwards, Pärn, Seo and Wong [38]		
Data-Driven	Lee, Lin, Seto and Migliaccio [39] Han, Lee and Peña-Mora [40] Chen, Qiu and Ahn [2] Alwasel, Elrayes, Abdel-Rahman and Haas [6] Yang, Wang and Chen [13]	IMU and vision sensors are used to collect data on whole-body or selected joints	The captured data are transformed into features, reflecting the characters or patterns of risky activities; classifiers can be trained through machine learning to identify interested activities. The classification rules are derived from data

application for outdoor construction practices, and more suitable for controlled experimental settings.

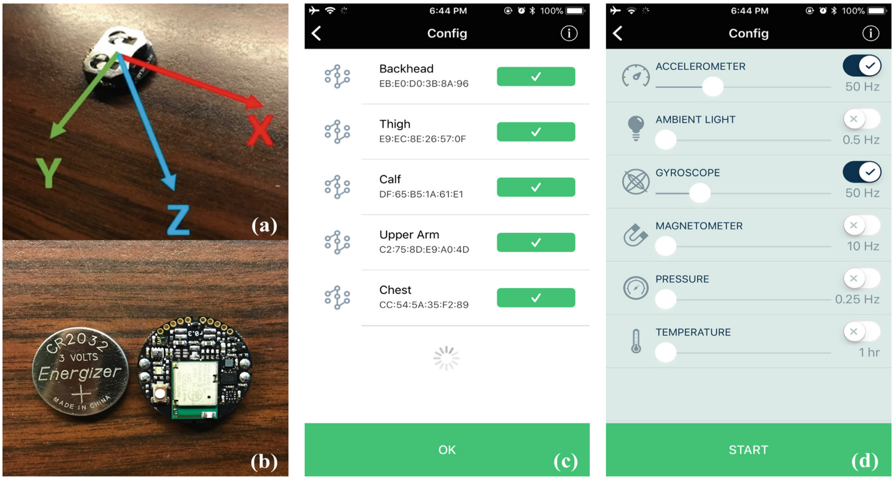
In contrast, a DD approach can potentially improve the practicability of sensing-based activity recognition. As no strict assumptions for sensor-to-segment orientation are needed, the DD approach focuses on the extracting features from general data patterns and training machine learning classifiers as activity recognition model. The sensors used are flexible, which can be low-cost IMU sensor [13]. The approach can base on the use of smart-phone built-in sensor [39]) or complex whole-body vision [40] and wearable IMU sensor system [2]. Both raw sensor output data and processed stable signals (such as Euler angle used in [2]) can be used for activity recognition model development. Clearly, the DD approach has the potential for activity recognition in construction activities [2, 28, 39].

There is an opportunity to develop a high-performance activity recognition model with reduced computational effort using low-cost wearable IMU sensors. Chen, Qiu and Ahn [2]'s described an example of recognizing the injury-prone activities in construction. Their approach is based on a whole-body IMU sensor placement approach for data collection. The raw data is transformed into high order tensors based on stable Euler angles to reduce computational cost. They achieved a high recognition performance for predefined injury-prone activities in two validations using support vector machine classifier. Some emerging work is focusing on reducing the computational effort through the selective sensor placement. This would accelerate the progress towards assessing activities in real-time, which is still an open research question in the area of sensor-based activity recognition [42]. Machine learning-based activity recognition can address some of these questions. However, for such an approach to work, there will be a need to further investigate the influence of critical success factors such as data segmentation window size, classification algorithm used, and feature set [13, 42]; the optimal model should be configured through an investigation of potential configuration combinations. There are variations triggered by the factor that the same construction activity is performed differently by different workers/same worker at different periods. Such variation can impede the scalable application and reusability of the DD approach and has resulted in a "cold-start" challenge in application [31]. Developing a generic rapid model configuration process could contribute to addressing such challenge.

3 Methodology

3.1 Hardware

A wearable data collection prototype was developed with IMU sensors to collect the motion data, each consists of triple-axis accelerometer, gyroscope, and magnetometer. Specifically, the IMU sensors used are the MetaMotion C sensors (by MBIENTLAB INC, San Francisco, CA), which record the data onto the Flash Memory on-board and also have a compared software, MetaBase, serving as the sensor controlling interface (Fig. 1c and d), formatting the logged data into CSV file, and exporting the data for analysis. The MetaBase allows connecting multiple sensors simultaneously, the total output frequency is limited up to 125 Hz. Noticeably, The output of magnetometer varies as the subjects orientation changes, the earth magnetic field might also confound the measurement [43]. These make it unsuitable for capturing human activities data with arbitrary orientation. In the tests, only accelerometer and gyroscope on the IMU sensor were used, both were set to max frequency of 50 Hz (as next level of frequency option is 100 Hz for each one, which will exceed the limit of 125 Hz in total for a sensor).



Backhead EB:E0:D0:3B:8A:96	✓
Thigh E9:EC:8E:26:57:0F	✓
Calf DF:65:B5:1A:61:E1	✓
Upper Arm C2:75:8D:E9:A0:4D	✓
Chest CC:54:5A:35:F2:89	✓

ACCELEROMETER	✓	50 Hz
AMBIENT LIGHT	×	0.5 Hz
GYROSCOPE	✓	50 Hz
MAGNETOMETER	×	10 Hz
PRESSURE	×	0.25 Hz
TEMPERATURE	×	1 hr

Fig. 1. (a) MetaMotion C IMU sensor board with the positive direction in each axis (b) Sensor Board and buttery (c) and (d) MetaBase for configuring sensors on a smartphone

3.2 Sensor Placement

Sensor placement has a significant impact on the activity recognition performance [42] and is related to the interested activities for recognition. One of the objectives here is investigating the proper sensor configuration to reduce the computational effort, sensors can be placed on the body parts corresponding to interested activities, such as a head sensor for neck movement, upper arm sensor for shoulder movement, chest center placement for upper body movement [44]. Sensors can be placed at thigh, calf, and ankle [23, 42] to track leg movements such as walking, and standing still, The focus here was limited to sensor placements on the thigh, calf, chest center, upper arm, and head. To further reduce the sensors used, the IMU sensors on leg and arm were attached to the right side for right-handed subjects, and five sensors were used in total. IMU sensors were fixed on the knee brace (thigh sensor and calf sensor), arm strap (upper arm sensor), chest center (chest sensor), and hardhat (head sensor) (Shown in Fig. 2) to ensure sensors tightly attached to the user and without attaching on the skin.

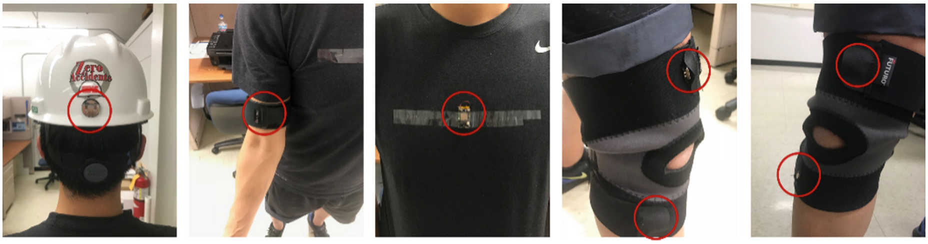


Fig. 2. Sensor placement (Head, Right arm, Chest, Right thigh, and Calf)

3.3 Activity Data Collection

Subject. The experiment was conducted on a healthy male graduate student (5′8′′, 161 lbs., right-handed), who has no history of bodily injury.

Training Experiment Data Collection. The targeted injury-prone activities used were identified from proceeding researches in [2, 10]. To mimic the scenarios where the recognition models will be applied, this research also included activities that are not of interest, such as standing still and walk, which are commonly seen in daily work [43]. Two sets of activities were conducted by the subject for data collection. The first set was related to the lower body, which included standing still, walk, squatting, kneeling, pick-up, and forward bending, as shown in Fig. 3.

The second set was related to the upper body, which included neck bending (forward bending), neck extension (backward bending), arm elevation, and working overhead, as shown in Fig. 4. Walking and standing still were also included in the second set.

The two set of activities were performed at regular pace respectively. To collect sufficient data for model training, each interested activity was performed ten times and holding for 2–3 s in that static state. the subject stood still for around 1 s between

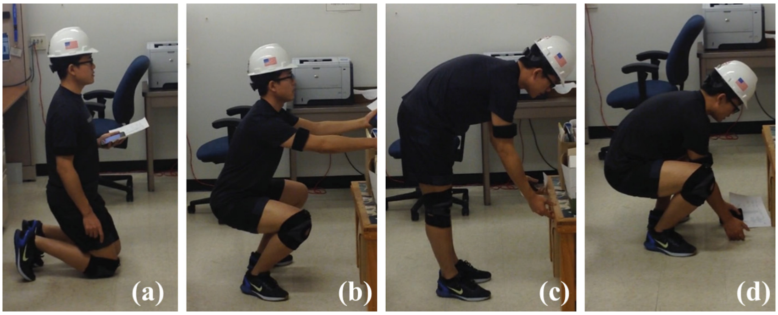


Fig. 3. Lower limb: (a) Kneeling (b) Squatting (c) Back bending (d) Pick up

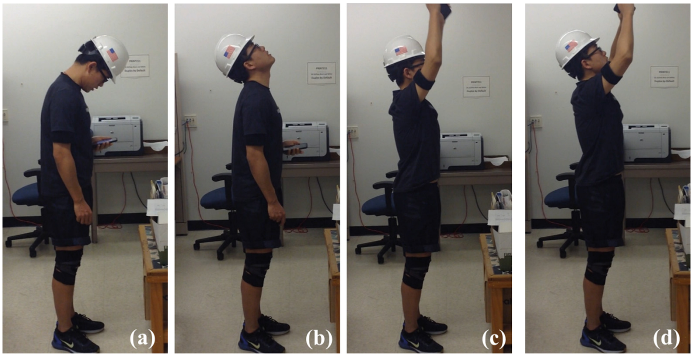


Fig. 4. Upper body: (a) Neck bending (b) Neck extension (c) Arm elevation (d) Work overhead

consecutive activities for separating consecutive activities. The training experiments with each set of activities were video-recorded as a ground truth for activities labeling. Although five sensors were used, this research focused more on the activity related sensors and continued the data analysis with those sensors corresponding to upper and lower body activities respectively to reduce the computation cost. The training experiment protocol is shown in the following Table 2.

Validation Experiment Data Collection. The training experiment data were used for developing classifiers. To validate the trained activity classifiers, another two sets of activities concerning lower and upper body were also performed by the subject. Each set of interested activities was performed for once in random order. The validation experiments were video-recorded for reference. The validation protocol is in Table 3.

3.4 Sensor Output Preprocessing and Labelling

The collected experiment data were temporarily stored on sensor boards, then processed “offline” by exporting data from each accelerometer and gyroscope in.CSV format.

Table 2. Training experiment data collection protocol

Activity sets	Interested sensors	Sequence of activities performed	Duration (approximately)	Sensor configuration
Lower body	Right thigh, Right calf, and Chest center	Standing still (ss)	5 s	50 Hz for accelerometer and gyroscope
		Walking (wk)	10 s	
		Kneeling (kn)	10 repetitions and holding 3 s	
		Squatting (sq)		
		Backbending (ba)		
		Pick up (pi)		
		Walking (wk)	10 s	
Upper body	Hardhat, Upper arm	Standing still (ss)[a]	5 s	
		Walking (wk)	10 s	
		Neck bending (nb)	10 repetitions with holding 3 s	
		Neck extension (ne)		
		Arm elevation (ae)		
		Work overhead (wo)		

[a] Waling and standing still were also included in the upper body activities although they are more likely to be considered as relating to lower body, this was done to mimic the normal daily work scenario as the interested activities might be minority during the work.

Synchronize the Output. Multiple sensors started and stopped together through the MetaBase, the ending time for each sensor still had a minor difference within one second. The effective data collection time was determined by the sensor with smallest ending time (dominating sensor, which had minimum data records). As the sensor output in the first second might be unstable, and that in the last second might have less than 50 records, the effective sensor operation time was set from the beginning of the 2nd second to the start of the last second of dominating sensor. The effective records number (N) from the dominating sensor was used to synchronize other sensors. The N consecutive data records from the beginning of 2nd second in each sensor were combined as a synchronized dataset, which dealt with the data loss (sometimes only 49 data points generated under 50 Hz).

Labelling and Windowing. The collected activity data in both training and validation experiments were manually labeled. This research also labeled the transition states to capture as much information of human activity as possible, which could potentially improve recognition performance [2]. To ensure the labeling reliability, the general data pattern was compared with video to identify the beginning and ending for each

Table 3. Validation experiment data collection protocol

Activity sets	Interested sensor	Sequence of activities performed	Duration (approximately)	Sensor configuration
Lower body	Right thigh, Right calf, and Chest center	Standing still (ss)	5 s	50 Hz for accelerometer and gyroscope
		Walking (wk)		
		Pick up (pi)	Conduct the activity in sequence in regular pace	
		Kneeling (kn)		
		Squatting (sq)		
		Backbending (ba)		
		Walking (wk)	5 s	
Upper body	Hardhat, Upper arm	Standing still (ss)	5 s	
		Walking (wk)		
		Work overhead (wo)	Conduct the activity in sequence in regular pace	
		Neck bending (nb)		
		Arm elevation (ae)		
		Neck extension (one)		
		Standing still (ss)	5 s	

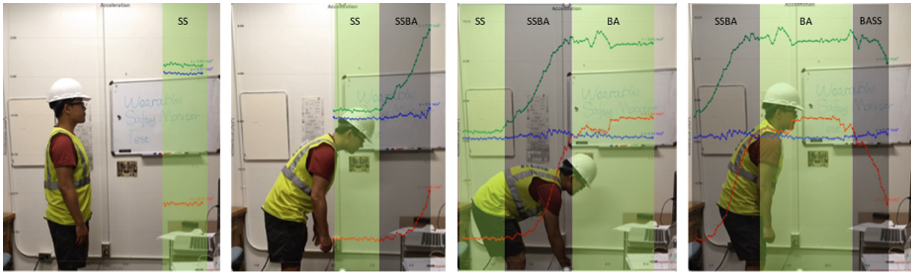


Fig. 5. Labeling activity of standing still (ss), standing still - backbending (ssba), backbending (ssba), and to backbending - standing still (bass)

activity and transition (Fig. 5). Instead of using a predefined fixed window, all data records falling into the specific interval of predefined activity will be labeled accordingly.

The optimal window size is not necessarily the same for interested activities varying in complexity and duration [23]. To investigate the proper window size, this research applied a step-wise search approach to testing the recognition performance of same classification algorithms under different window sizes. An automatic "majority vote" was adopted to label the activity for a resized window, specifically, the label of majority records in a window will be the label of the entire window, which is convenient for automatically relabeling with different window sizes.

The window sizes for daily activities can range from 0 to 7 s, and most of the reviewed studies adopted the interval of 0–1 s and 2–3 s [23]. Through checking motion data collected in the experiments, it was found that the interested activities in this research did not exceed 3 s, therefore, an interval of 0.5 s to 2.5 s was used for searching optimal window size, the search step was set as 0.1 s. Besides, a 50% overlap between adjacent windows has been repeatedly suggested to enhance human activities recognition [22, 45], such window overlap was also implemented in this research.

3.5 Features Extraction

Full Feature Construction. After segmenting activity data stream into consecutive labeled windows, features can be extracted to characterize the activity in the window. Various features can be extracted from the continuous time-related sensor signals, such as time-domain, frequency-domain, heuristic, and domain-specific features. This research used features in (1) time-domain, which describes the basic waveform characteristics and signal statistics; and (2) frequency-domain, which focuses on the periodic structure of the signal [5]. For frequency-domain features, the Fourier Transform (FT) was applied to the raw sensor output to acquire the estimated spectral density of the time series. The output of the transform is a signal frequency spectral, which includes the information about signals in different frequencies within a time series data and corresponding amplitudes of the signal. Frequency-related features can be extracted after the transform.

The feature extraction was implemented with corresponding packages in R (3.4.1)[1], a description of the features used and calculation method is shown in Table 4.

Spectral was used for calculating all the other frequency-domain features and would not be used for training recognition models. 13 types of features were extracted from 3 axes of the accelerometer and gyroscope on each sensor. As five IMU sensors were used, for each window, a total of 390 features (5 IMU × 2 sensor units on IMU × 3 axes × 13 features types) could be extracted from each window.

Features Selection. Features might be irrelevant and barely contribute to improving model performance when the feature space dimension is too high [13]. A high dimensional feature space also increases the computation cost and memory requirement, particularly when the activity recognition model will be applied in real-time. One task of this research is to find an optimal subset of full feature to reduce the

[1] https://www.r-project.org/about.html.

Table 4. Features extraction

Category	Types	Description	Package
Time-Domain	Minimum	Basic statistics of sensor output in a given window	Stats (3.4.1)
	Maximum		
	Mean		
	Variance		
	Average Absolute Deviation	Mean absolute deviations from a central point	
	Slope	Sen's slope for a series of data	Trend (1.0.1)
	Root Mean Square	Square root of arithmetic mean	Seewave (2.0.5)
Frequency-Domain	Spectral	Frequency spectral after FT	
	Entropy	Shannon entropy	
	Spectral centroid	Centroid of a given spectrum	
	Skewness	Symmetric of distribution	
	Kurtosis	Heavy tail of distribution	
	Signal energy	Sum squared signal amplitude [1]	
	Frequency range power	Sum of absolute amplitude of the signal [3]	

computational effort and maintain the acceptable model performance. Optimal subsets characterizing interesting activities will be investigated with feature selection.

Two approaches could be applied to reduce the feature dimension. One is extracting the commonalities and reducing features with high correlation, such as the Principal Component Analysis (PCA). Another is selecting the “best” subset giving the strongest discrimination power. In comparison, the second needs more computational effort, but will always find the “best” feature combination for recognition.

This research aims at exploring the subsets to improve recognition performance and adopted the second approach. Specifically, the Recursive Feature Elimination (RFE) method was used, which uses an optimization algorithm to find the best feature set through creating models and keeping aside the best or the worst performing features iteratively. It constructs the next model with left features until all features are exhausted and stores best feature as a best-feature set. The REF process was implemented using Caret (version 6.0-77) package in R with Random Forest to evaluate models.

Features Transform. As features have different units and scale, all features were normalized by reducing the mean then dividing their standard deviation before training the full-featured model. Features were also normalized before applying the RFE.

3.6 Activity Recognition Model Selection

Classification Algorithms. Extracted feature sets were used as input for model training. Multiple classification algorithms have shown promising performance for daily activities recognition, such as threshold method, decision tree (C4.5), Naïve Bayes (NB) and K-Nearest Neighbor (KNN) [23, 42]. Particularly, Support Vector Machine (SVM) has shown great performance in recognizing construction workers' awkward postures [2]. This research investigated the performance of four commonly used recognition algorithms under different window sizes, namely NB, KNN, SVM, and C4.5.

Model Training

Training Procedure. The experiment datasets were shuffled and randomly split into 80% for training and 20% for testing. A fixed random seed was used when splitting the data to ensure the repeatable result. The four classifiers were trained on the same training set. In order to further improve the recognition performance, a parameter "tuning" process was implemented for each classifier for different window sizes and activities. Performance of the same classifier with varying parameters was compared through 10-fold cross-validation in training. The tuned classifier was validated on 20% testing set.

Parameter Tuning. The parameter tuning process is shown in Table 5. For the Naïve Bayes model, default setup in the R package e1071(Version 1.6-8) was used.

Table 5. Parameter setup and tuning

Algorithm	Parameters	Tuning process
SVM	Epsilon in insensitive-loss function	Searching from 0 to 1 by steps of 0.1
	Cost of constraints violation	1 by default
	Kernel	Radial by default
KNN	K: Number of neighbors	Searching from 1 to 20 by steps of 1
C4.5	Complexity Parameter (CP), any split that does not decrease the overall lack of fit by cp is not attempted	Searching the best CP value automatically in decreasing sequence until reaching an extremely small value, the CP value gives minimum cross-validation error will be used for pruning

Model Selection and Validation

Model Performance Comparison. The performance of each tuned "best" model from the same training dataset was compared to the same testing set, then evaluated for overall classification performance and computational efforts. For the performance, accuracy and kappa value (reflecting influence from the imbalanced distribution of activity types) were compared. The model training and operation time for prediction was evaluated for computational efficiency.

The tuned model's performance under each window size was compared. The window size giving an overall high performance among models was selected for specific activities. Model with an overall best performance in accuracy and computational efficiency was selected as the "best" model under the selected best window size.

Model Performance Validation. The validation experiment dataset was windowed with selected best window size to validate the configured model. The identified best model was re-trained with the entire training experiment dataset, then the reconfigured best model's performance would be tested on the validation experiment dataset.

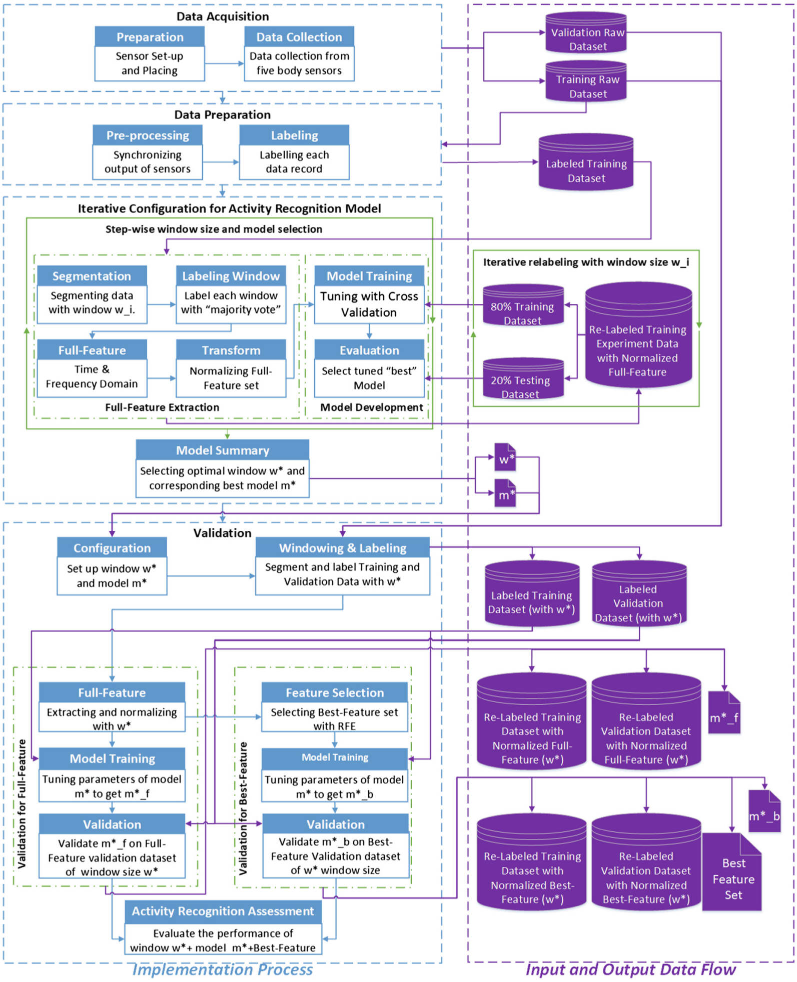

Data Acquisition
Preparation
Sensor Set-up and Placing
Data Collection
Data collection from five body sensors
Validation Raw Dataset
Training Raw Dataset
Data Preparation
Pre-processing
Synchronizing output of sensors
Labeling
Labelling each data record
Labeled Training Dataset
Iterative Configuration for Activity Recognition Model
Step-wise window size and model selection
Segmentation
Segmenting data with window w_i.
Labeling Window
Label each window with "majority vote"
Model Training
Tuning with Cross Validation
Full-Feature
Time & Frequency Domain
Transform
Normalizing Full-Feature set
Evaluation
Select tuned "best" Model
Full-Feature Extraction
Model Development
Iterative relabeling with window size w_i
80% Training Dataset
20% Testing Dataset
Re-Labeled Training Experiment Data with Normalized Full-Feature
Model Summary
Selecting optimal window w* and corresponding best model m*
w*
m*
Validation
Configuration
Set up window w* and model m*
Windowing & Labeling
Segment and label Training and Validation Data with w*
Labeled Training Dataset (with w*)
Labeled Validation Dataset (with w*)
Full-Feature
Extracting and normalizing with w*
Model Training
Tuning parameters of model m* to get m*_f
Validation
Validate m*_f on Full-Feature validation dataset of window size w*
Validation for Full-Feature
Feature Selection
Selecting Best-Feature set with RFE
Model Training
Tuning parameters of model m* to get m*_b
Validation
Validate m*_b on Best-Feature Validation dataset of w* window size
Validation for Best-Feature
Re-Labeled Training Dataset with Normalized Full-Feature (w*)
Re-Labeled Validation Dataset with Normalized Full-Feature (w*)
m*_f
m*_b
Re-Labeled Training Dataset with Normalized Best-Feature (w*)
Re-Labeled Validation Dataset with Normalized Best-Feature (w*)
Best Feature Set
Activity Recognition Assessment
Evaluate the performance of window w*+ model m*+Best-Feature
Implementation Process
Input and Output Data Flow

Fig. 6. Process map of activity recognition implementation

Validation with the Best Features. The full-feature was substituted with best features in both experiments and validation to re-train the best model with selected window size. The performance was compared to assess the selected best features.

The data collection, processing, model configuration process described is organized in Fig. 6, which is coded and implemented in R.

4 Result and Discussion

4.1 Data Description

The data collected in the training experiments and validation experiments were pre-processed and labeled, results were summarized in the following Tables 6 and 7.

Table 6. Raw data sets

		Sensors	Operation time (s)	Max difference (s)	Records in raw data	Effective records
Training	Lower body	Thigh	281.07	0.09	14112	14055
		Calf	280.98		14256	
		Chest center	281.01		14208	
	Upper body	Hardhat	233.25	0.18	11521	11454
		Upper arm	233.07		11640	
Validation	Lower body	Thigh	51.24	0.15	2574	2511
		Calf	51.09		2595	
		Chest center	51.15		2589	
	Upper body	Hardhat	40.35	0.01	1995	1926
		Upper arm	40.36		2016	

From the Table 6, the difference between sensors' operation time was within 0.2 s, which was relatively minor compared to the sensor operation time (no more than 0.3%). The manual synchronization process could be applied. For the labeled data in Table 7, the majority of the activities in experiments (over 60%) are standing still, walking, and transition states. Such result resembles the real-world situation, where targeted activities are mixed with a large proportion of not-interested daily activities [43]. Such imbalance poses a challenge for recognition, while it also helps to validate the performance of proposed approach.

Table 7. Labeled activities

	Activity label	Training experiment		Validation experiment	
		Records	Percentage	Records	Percentage
Lower body	Standing still (ss)	3537	25.42%	743	29.59%
	Walking (wk)	1172	8.42%	649	25.85%
	Kneeling (kn)	1259	9.05%	138	5.50%
	Squatting (sq)	1210	8.70%	83	3.31%
	Backbending (ba)	839	6.03%	114	4.54%
	Pick up (pi)	1212	8.71%	151	6.01%
	Transition	4683	33.66%	633	25.21%
Upper body	Standing still (ss)	3092	26.99%	760	39.46%
	Walking (wk)	531	4.64%	418	21.70%
	Neck bending (nb)	818	7.14%	87	4.52%
	Neck extension (ne)	952	8.31%	68	3.53%
	Arm elevation (ae)	997	8.70%	79	4.10%
	Working overhead (wo)	941	8.22%	115	5.97%
	Transition	4123	36.00%	399	20.72%

4.2 Searching for Optimal Window Size and Model

The proposed step-wise optimal window size searching process was implemented with R (3.4.1) on a Windows 10 PC (Intel Core i7-7700 CPU@ 2.8 GHz, 16 GB RAM). Under each window (w_i), the step-wise parameter tuning process was implemented for classification algorithms. The window searching result is shown in the following Figs. 7 and 8. The step-wise searching result serves as the information needed to select

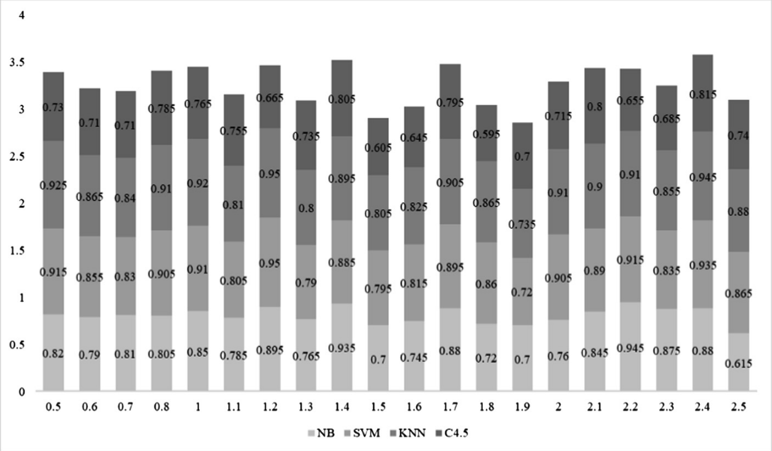


Fig. 7. Performance in lower body activity in training experiment data set

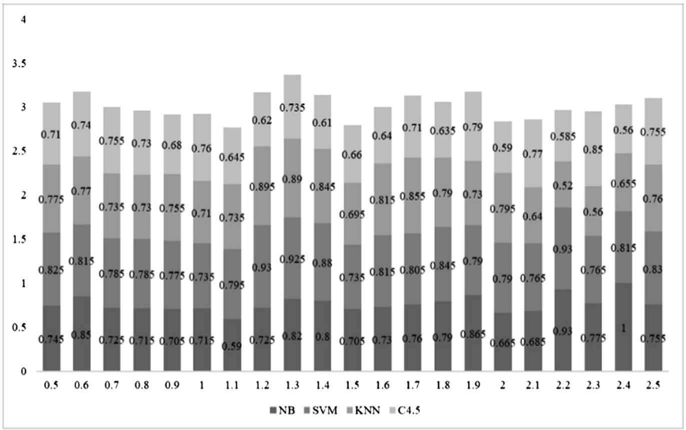


Fig. 8. Performance in upper body activity in training experimental data set

the optimal window size and corresponding model, users can develop the customized evaluation rules to meet requirements, e.g. computational effort or accuracy. This research adopted the following approach as an example of selecting the window size and model.

Evaluation Metrics of Activity Recognition Models

Computational Efficiency. The computational efficiency was evaluated through model training and operation time for classification. The increased window size leads to decreased window number, and also a general trend of decreasing model training time. There was a great variation for the training time under same window size. Besides the computational complexity difference of classification algorithms used, another reason for the great variation is the difference in parameter tuning process. The process determined the number of iterations needed to find the best parameter combination. NB model had the shortest training time as no tuning process needed; while KNN model had the longest training time due to a 20-iteration used for finding optimal K under each window. The classification time was very short with the trained model and could be negligible and identified as zero by the program.

Activity Recognition Performance. Both accuracy and kappa coefficients were considered through a relative crude weighted-average approach and given equal importance by assigning a weight of 0.5 for each. The average value of accuracy and kappa was constructed as a temporary model performance metric. The performance with such metric of models under different window sizes is shown in Figs. 7 and 8.

Overall Evaluation. The model training time was not compared directly as it is influenced by the predefined parameter tuning process. The operation time of all models was relatively short in the test, the classification computation efficiency of models in same window size could be treated as equal. In this case, the average of accuracy and kappa as an overall metric for recognition performance.

Recognition Model Configuration Result. This research considered the overall performances of multiple classification algorithms under given window size to determine the preferred window size. Specifically, when the average of accuracy and kappa was more than 0.9, it was treated as a "good" model, the window size gave maximum good models were selected. If the number of good models was same, this research selected the window with higher average accuracy and kappa. Window size 1.2s and SVM were finally selected for lower body activities; 2.2s and NB were selected for upper body activities.

Based on the results in Figs. 7 and 8, there was no monotonic relationship between recognition performance and window size, the widely used widow sizes 0.5s or 2.5s did not outperform other window sizes. The selected window sizes in this research are in line with the that for fitness activities [23], which concluded a range of 1–2s window provided the best trade-off between recognition speed and accuracy for on-body activity recognition system. Although the actives were performed by the same subject within 90 min, the optimal window size varies between activities, which makes it more meaningful to investigate an optimal window size for a specific activity.

SVM was identified to outperform other models in terms of recognition performance among the four classification algorithms applied. The high performance of SVM was also reported in [2], which was applied with tensor decomposition and achieved a recognition accuracy of 63% and 58% in validation. The SVM model is a preferable method for processing data with high feature space dimension, even for data sets with a feature dimension more than the sample size. In this research, the activity data can have a feature space as high as 235, which is relatively high even for the most frequent activity type (standing still) and may exceed the records of some activities in validation experiments. The SVM model is preferable in this situation.

4.3 Feature Selection

A "greedy" RFE approach was applied on Full-Feature (FF) set to select the best subset as the Best-Feature (BF). An examination of selected features is shown below (Table 8).

Table 8. Feature selection result

	Window (s)	FF size	BF size	RFE time (s)	Examination of best-feature[a]		
					Sensor placement	Category	Sensor units
Lower body	1.2	235	130	1169.38	Calf	Time domain	Accelerometer
Upper body	1.3[b]	157	33	422.12	Upper arm		

[a] Most frequently used features regarding sensor placement, feature category, and sensor units.
[b] 2.2s was dropped, which will be discussed later in Sect. 4.4.

The RFE process selected 55% and 21% of the full features to be the best features for upper body and lower body activities respectively. The exhaustive feature selection approach is computation intensive, which should be implemented “offline”. Assessment of the activity recognition performance with selected features will be provided in Sect. 4.4.

The feature selection result can also throw light upon what kind of features contribute more for recognizing construction workers’ activities. Firstly, features from accelerometer contribute more than those from gyroscope in this research. This might be attributed to the nature of activities in the experiment. Construction activities performed are not highly fierce and intensive like those in the sports, where the gyroscope might be preferable to characterize the energetic activities. Secondly, the nature of activities may also explain that most of the features selected are Time-Domain features, which reflect the general waveform of the sensor signal. Frequency Domain features are more commonly used to characterize activities with different complexity [45]. Finally, the calf was identified as more useful for characterizing lower body activities than thigh, future tests can be conducted to explore whether the sensors required could be reduced to improve computation efficiency.

4.4 Validation of Proposed Approach

The validation was performed by re-windowing the training experiment dataset with selected window, then training the model with the selected algorithm. The trained model was then used to recognize activities in the re-windowed validation experiment dataset. The validation process was performed under both Full-Feature and Best-Feature to evaluate the feature selection performance. The detailed optimal window size and model searching process result in training experiment data was stored[2], which would be used for assessing and selecting the best configuration.

Noticeably, when a 2.2s window size was used for segmenting upper body validation data set, the Arm Elevation (AE) and Neck Extension (NE) did not show up in the data set. Examination of validation experiment data showed that there were only 68 records of NE and 79 records of AE collected under 50 Hz frequency, which is less than the 2.2s window size. When mixed with other activities in the same window, such activities tend to be the minority and represented by the label of majority activity in the same window. Therefore, the 2.2s window size was dropped, the next “best” window size (1.3s) was searched and adopted to avoid missing of activities in the validation experiment.

Feature Selection Assessment. From the result in Table 9, the best subset significantly decreased the model training time in two sets of validation experiments. When classifying new record, although the SVM model with selected features only decreased the operation time by 0.03 s, such reduction in latency will be helpful towards potential real-time application in mobile devices, which are not equipped with high-performance processors as computers used in this research.

[2] The experiment result is not attached in the paper due to limited spaces, the authors can provide the original result upon request.

Table 9. Comparison of validation result

Validation experiment	Window size and model	Windows	Training time (s)				Operation time (s)		Accuracy	
			Kappa							
FF	BF	FF	BF	FF	BF	FF	BF			
Lower body	1.2s+SVM	43	12.82	7.55	0.01	0.0	0.72	0.81	0.65	0.77
Upper body	1.3s+SVM	31	6.98	1.93	0.03	0.0	0.74	0.74	0.67	0.67

In terms of the recognition performance, the overall performance did not decrease compared to full features and even increased by 12.5% for the lower body recognition. One reason might be the dropped features caused overfitting problem in the model training. As these dropped features did not characterize the activities as well as selected features, the over-trained model on full features would not perform well on new records.

The overall assessment on selected best feature set has shown an improvement in the recognition performance considering both efficiency and accuracy. The identified best feature set could also be reused for recognizing the same set of activities, as such activities are well characterized by the selected features.

Activity Recognition Assessment. A close examination of the recognition performance for each performed activity is shown in Table 10.

The sensitivity, also known as recall, measures the proportion of positives that are correctly identified as such; the specificity measures the proportion of negatives that are correctly identified as such. As the interested activities were a minority, prevalence (proportion) of activities was also provided for reference. The balanced accuracy is a preferred metric to assess the model performance, and calculated as following [46]:

$$Balanced\ Accuracy = \frac{1}{2} \times \left(\frac{True\ Positive}{Positve} + \frac{True\ Negative}{Negative} \right) \tag{1}$$

The activities in both validation experiments were generally grouped as interested, transition, and normal activities. For the lower body activities, the proposed approach gave an acceptable result. The overall performance of the interested activities recognition reached an accuracy of 0.93 on average; the normal activities recognition accuracy also reached 0.885. The transition states were identified separately, the balanced accuracy for recognizing transition states was lowest (0.81 on average). Through exanimating confusion matrix for transitions states SSBA, PISS, and SQSS, they were classified as SS, SS or PI, and PI respectively. The transition states might be confused with their adjacent activities in sequential activities, as the beginning and end of a transition are similar to their preceding and subsequent activities. This could also be caused by the researchers' subjective bias in manual labeling process. The transition states tend to be confounded with adjacent activities when they start and end. For the misclassification of SQSS and PI, it might be due to the similarity between pick up (squatting and back bending) and squatting activities. Although recognition performance for transition states was low, separating these activities has been suggested for improving the recognition of interested activities [2], and shown a promising result this research.

Table 10. Validation result for activities

			Sensitivity		Specificity		Prevalence	Balanced accuracy	
			FF	BF	FF	BF	FF&BF	FF	BF
Lower body activities in validation experiment	Interested activities	BA	0.50	0.50	1.00	1.00	0.05	0.75	0.75
		KN	1.00	1.00	1.00	1.00	0.05	1.00	1.00
		PI	0.00	1.00	0.93	0.95	0.05	0.46	0.98
		SQ	1.00	1.00	0.95	1.00	0.05	0.98	1.00
	Transition	SSBA	0.00	0.00	0.98	0.98	0.02	0.49	0.49
		SSKN	1.00	1.00	1.00	1.00	0.05	1.00	1.00
		BASS	1.00	1.00	1.00	1.00	0.02	1.00	1.00
		KNSS	1.00	1.00	1.00	1.00	0.02	1.00	1.00
		PISS	0.00	0.00	1.00	1.00	0.05	0.50	0.50
		SQSS	0.00	0.00	1.00	1.00	0.02	0.50	0.50
		SSPI	0.00	1.00	1.00	1.00	0.02	0.50	1.00
		SSSQ	1.00	1.00	1.00	1.00	0.02	1.00	1.00
	Normal activities	SS	1.00	1.00	0.79	0.83	0.33	0.90	0.91
		WK	0.64	0.73	1.00	1.00	0.26	0.82	0.86
Upper body activities in validation experiment	Interested activities	AE	1.00	0.00	0.97	0.97	0.03	0.98	0.48
		NB	1.00	1.00	0.97	0.97	0.03	0.98	0.98
		NE	1.00	1.00	1.00	1.00	0.03	1.00	1.00
		WO	0.00	0.00	0.97	0.93	0.03	0.48	0.47
	Transition	SSAE	0.00	1.00	1.00	1.00	0.03	0.50	1.00
		SSNB	1.00	1.00	1.00	1.00	0.03	1.00	1.00
		AESS	0.00	0.00	1.00	1.00	0.06	0.50	0.50
		NBSS	0.00	0.00	1.00	1.00	0.03	0.50	0.50
		NESS	1.00	1.00	1.00	1.00	0.03	1.00	1.00
		SSNE	1.00	1.00	1.00	1.00	0.03	1.00	1.00
		SSWO	1.00	1.00	1.00	0.97	0.03	1.00	0.98
		WOSS	1.00	0.00	1.00	1.00	0.03	1.00	0.50
	Normal activities	SS	1.00	1.00	0.75	0.85	0.35	0.88	0.93
		WK	0.57	0.71	1.00	1.00	0.23	0.79	0.86

For the lower body activities, the AE and WO were misclassified with each other. As these two activities were only different with whether the subject had a neck extension, the combination of head and arm sensor placement may not be sufficient to differentiate them. The confusion matrix also showed that the misclassification for transition states happened between their adjacent activities as in the lower body validation experiment. Noticeably, although the feature selection in this validation maintains the same overall accuracy, the performance of recognizing interested activities decreased. The increased performance happened for transition and normal activities, which were not of research interest. This can be caused by the object function of RFE's optimization process, as it uses the overall recognition accuracy of all activities. Features dropped under such object might not be useful for improving the recognition of interested activities. In this case, one way to further improve the feature selection performance is modifying the object function of RFE to maximize the recognition performance for interested activities.

Rapid Model Configuration Process. This research adopted a selective sensor placement approach to identify targeted injury-prone activities. The subsequent module configuration process and performance result under different configuration can be provided (as shown in Figs. 7 and 8), where user-defined rules should be applied to finalize the configuration. Once the user determined the sensor placement, selection criteria for configured models, the proposed process in Fig. 6 can be developed into an integrated rapid automatic configuring process to customize Data-Driven recognition models.

This research selected the sensor placements on activity related body parts and gave equal importance for model accuracy and efficiency in configuration. The proposed rapid model configuration process has shown a promising result in validation, particularly for the interested lower body activities, which gave an average balanced accuracy of 0.93, and also an operation time less than 0.01 s. For the upper body activities, the chosen two-sensor placement might be insufficient to differentiate similar activities, more sensor placement plans can be investigated under the proposed approach. In terms of model development, an exhaustive searing process was adopted to configure an optimal recognition model and used an average of 9.9 s in configuring two models. The model development time can be reduced by 52% to an average of 4.74 s through feature selection, which was a significant reduction in computational efforts.

The abovementioned process gives an example of how the recognition model was configured with user-defined rules. Various parameters can be modified and investigated to meet users' requirements. When the configuration rules are specified, the proposed generic process for rapid and automatic configuration can be reused for deploying customized model for varying users and their activities. Users' activities may vary among individuals, the proposed process can be adopted rapidly to build a high-performance model for improving the practicality, scalability, and reusability of the Data-Driven approach.

5 Conclusion, Limitation, and Future Work

5.1 Conclusion

Injury prevention through activity recognition has been suggested as an effective approach for preventing accumulative injuries in construction. This research is directed at improving the performance and practicality of the Data-Driven activity recognition approach. The selective body sensor placement, step-wise model configuration, and "greedy" feature selection approaches were investigated to improve the recognition accuracy and reduce the computational effort in this research. The feasibility of the proposed approach was assessed through two sets of experiments. Based on the investigation, the following conclusions can be made.

Selective Sensor Placement. From the assessment of the selective sensor placement, the thigh and calf are suitable placements for recognizing the lower body activities has shown relatively high performance. More tests can be done to investigate the optimal upper body sensor placement. To conclude the selective placement can be adopted to achieve high recognition performance with the reduced computational effort for targeted pre-defined injury-prone activities in construction.

Window Size. This research found the optimal window sizes of 1.2s and 1.3s for the two sets of activities performed by the same subject, which is different from the widely used 0.5s or 2.5s in previous research. It is preferable to select the window size for specific activities in a customized approach instead of adopting a fixed window size. Noticeably, with the "majority vote" approach for labeling windows, the window size should not be too large comparing to the activity duration, in case of failing to identify activities happened in short time.

Feature Selection. Feature selection can improve the accuracy and reduce computational efforts of activity recognition. While in terms of the injury-prone activities interested, the performance improvement is not consistent between two sets of activities. The feature selection criteria in the greedy optimization algorithm of RFE are a major reason for such result as discussed in foregoing section. A more effective feature selection under RFE can be investigated by modifying the object function as maximizing the recognition accuracy of targeted activities.

The examination of the feature selection result shows that features from accelerometer contributed more for recognition comparing to the gyroscope. The features reflecting the sensing data waveform contributed more compared to the frequency domain features for the activities in experiments. These findings can provide insights for feature construction when developing machine learning classifiers for similar activities.

Classification Algorithm Selection. SVM outperformed the other three models used in both sets of activities in terms of accuracy. The model training time is influenced by the parameter tuning process specified, while the model operation time for all the four models used was relatively short (less than 0.4 s), and SVM's operation time for classifying new data record is less than 0.01 s with selected features. In conclusion, SVM is a preferable model for classifying the interested injury-prone activities.

In summary, the proposed Data-Driven approach can be implemented successfully with user-determined sensor placement and model selection rules. The configured activity recognition model has shown acceptable results with an overall accuracy of 0.81 and 0.74 respectively for each set of validation activities. Both the model developing and operation time were also reduced significantly through proposed feature selection method. The proposed approach can be developed into the automatic configuring process with user-defined sensor placement and model selection criteria. The proposed approach and findings from this research are useful for not only the sake of research but for the system design purposes or design task of injury prevention in other areas like sports or manufacturing.

5.2 Limitation and Future Work

The scope of work directing at investigating the proposed approach has been specified in the preceding sections, the following aspects were identified for improvement, and will be explored in the further work.

- Sensor Synchronization. The sensor output was synchronized manually, future exploration is needed for synchronizing the data more precisely.
- Data Labeling. A post-labeling process with video reference is time-consuming and sometimes inaccurate, particularly for the transitions between activities. The further test could try adopting the subject's self-reporting approach to determine the starting and ending of activities, which is more efficient for pre-processing the data.
- Duration of Validation Experiment. The duration of validation experiment was relatively short in this research. As the IMU sensor might suffer from drift in long time operation, long-term experiments should be conducted for assessing the sensor output quality.
- Validation with More Subjects. The proposed process was implemented on a single subject in this research, the future test will be done with multiple subjects to validate the rapid configuration approach.

References

1. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data (2005)
2. Chen, J., Qiu, J., Ahn, C.: Construction worker's awkward posture recognition through supervised motion tensor decomposition. Autom. Constr. **77**, 67–81 (2017)
3. Heinz, E.A., Kunze, K.S., Gruber, M., Bannach, D., Lukowicz, P.: Using wearable sensors for real-time recognition tasks in games of martial arts-an initial experiment. In: 2006 IEEE Symposium on Computational Intelligence and Games, pp. 98–102. IEEE (2006)
4. Rosenman, K.D., Kalush, A., Reilly, M.J., Gardiner, J.C., Reeves, M., Luo, Z.: How much work-related injury and illness is missed by the current national surveillance system? J. Occup. Environ. Med. **48**, 357–365 (2006)
5. Yang, G.-Z., Yang, G.: Body Sensor Networks. Springer, London (2006). https://doi.org/10.1007/1-84628-484-8
6. Alwasel, A., Elrayes, K., Abdel-Rahman, E., Haas, C.: A human body posture sensor for monitoring and diagnosing MSD risk factors. FUTURE (2013)
7. Rwamamara, R., Lagerkvist, O., Olofsson, T., Johansson, B., Kaminskas, K.A.: Prevention of work-related musculoskeletal injuries in construction industry (2010)
8. Kim, H., Ahn, C.R., Engelhaupt, D., Lee, S.: Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement. Autom. Constr. **87**, 225–234 (2018)
9. Li, G., Buckle, P.: Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods. Ergonomics **42**, 674–695 (1999)
10. Valero, E., Sivanathan, A., Bosché, F., Abdel-Wahab, M.: Analysis of construction trade worker body motions using a wearable and wireless motion sensor network. Autom. Constr. **83**, 48–55 (2017)
11. Yan, X., Li, H., Li, A.R., Zhang, H.: Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. Autom. Constr. **74**, 2–11 (2017)
12. Lavender, S., Li, Y., Andersson, G., Natarajan, R.: The effects of lifting speed on the peak external forward bending, lateral bending, and twisting spine moments. Ergonomics **42**, 111–125 (1999)

13. Yang, J.-Y., Wang, J.-S., Chen, Y.-P.: Using acceleration measurements for activity recognition: an effective learning algorithm for constructing neural classifiers. Pattern Recogn. Lett. **29**, 2213–2220 (2008)
14. Wang, D., Dai, F., Ning, X.: Risk assessment of work-related musculoskeletal disorders in construction: state-of-the-art review. J. Constr. Eng. Manag. **141**, 04015008 (2015)
15. Han, S., Lee, S.: A vision-based motion capture and recognition framework for behavior-based safety management. Autom. Constr. **35**, 131–141 (2013)
16. Kim, H., Ahn, C.R., Stentz, T.L., Jebelli, H.: Assessing the effects of slippery steel beam coatings to ironworkers' gait stability. Appl. Ergon. **68**, 72–79 (2018)
17. Seel, T., Schauer, T., Raisch, J.: Joint axis and position estimation from inertial measurement data by exploiting kinematic constraints. In: 2012 IEEE International Conference on Control Applications (CCA), pp. 45–49. IEEE (2012)
18. Seel, T., Raisch, J., Schauer, T.: IMU-based joint angle measurement for gait analysis. Sensors **14**, 6891–6909 (2014)
19. Maman, Z.S., Yazdi, M.A.A., Cavuoto, L.A., Megahed, F.M.: A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. Appl. Ergon. **65**, 515–529 (2017)
20. Golabchi, A., Han, S., Seo, J., Han, S., Lee, S., Al-Hussein, M.: An automated biomechanical simulation approach to ergonomic job analysis for workplace design. J. Constr. Eng. Manag. **141**, 04015020 (2015)
21. Akhavian, R., Behzadan, A.H.: Smartphone-based construction workers' activity recognition and classification. Autom. Constr. **71**, 198–209 (2016)
22. Joshua, L., Varghese, K.: Accelerometer-based activity recognition in construction. J. Comput. Civ. Eng. **25**, 370–379 (2010)
23. Banos, O., Galvez, J.-M., Damas, M., Pomares, H., Rojas, I.: Window size impact in human activity recognition. Sensors **14**, 6474–6499 (2014)
24. Wang, D., Dai, F., Ning, X., Dong, R.G., Wu, J.Z.: Assessing work-related risk factors on low back disorders among roofing workers. J. Constr. Eng. Manag. **143**, 04017026 (2017)
25. Gruetzemacher, R., Gupta, A., Wilkerson, G.B.: Sports injury prevention screen (SIPS): design and architecture of an Internet of Things (IoT) based analytics health app. In: CONF-IRM, p. 18 (2016)
26. Valero, E., Sivanathan, A., Bosché, F., Abdel-Wahab, M.: Musculoskeletal disorders in construction: a review and a novel system for activity tracking with body area network. Appl. Ergon. **54**, 120–130 (2016)
27. Chen, J., Ahn, C., Han, S.: Detecting the hazards of lifting and carrying in construction through a coupled 3D sensing and IMUs sensing system. In: 2014 International Conference for Computing in Civil and Building Engineering (2014)
28. Jebelli, H., Ahn, C.R., Stentz, T.L.: The validation of gait-stability metrics to assess construction workers' fall risk. In: American Society of Civil Engineers (ASCE) (2014)
29. Bartalesi, R., Lorussi, F., Tesconi, M., Tognetti, A., Zupone, G., De Rossi, D.: Wearable kinesthetic system for capturing and classifying upper limb gesture. In: First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics Conference, World Haptics 2005, pp. 535–536. IEEE (2005)
30. Yang, K., Aria, S., Ahn, C.R., Stentz, T.L.: Automated detection of near-miss fall incidents in iron workers using inertial measurement units. In: Construction Research Congress, pp. 935–944 (2014)
31. Chen, L., Hoey, J., Nugent, C.D., Cook, D.J., Yu, Z.: Sensor-based activity recognition. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**, 790–808 (2012)

32. Golabchi, A., Han, S., Fayek, A.R., AbouRizk, S.: Stochastic modeling for assessment of human perception and motion sensing errors in ergonomic analysis. J. Comput. Civ. Eng. **31**, 04017010 (2017)
33. Keyserling, W.M., Brouwer, M., Silverstein, B.A.: The effectiveness of a joint labor-management program in controlling awkward postures of the trunk, neck, and shoulders: results of a field study. Int. J. Ind. Ergon. **11**, 51–65 (1993)
34. O'brien, R.M.: A caution regarding rules of thumb for variance inflation factors. Qual. Quant. **41**, 673–690 (2007)
35. Fang, Y., Cho, Y.K., Druso, F., Seo, J.: Assessment of operator's situation awareness for smart operation of mobile cranes. Autom. Constr. **85**, 65–75 (2018)
36. Alwasel, A., Sabet, A., Nahangi, M., Haas, C.T., Abdel-Rahman, E.: Identifying poses of safe and productive masons using machine learning. Autom. Constr. **84**, 345–355 (2017)
37. Wang, D., Chen, J., Zhao, D., Dai, F., Zheng, C., Wu, X.: Monitoring workers' attention and vigilance in construction activities through a wireless and wearable electroencephalography system. Autom. Constr. **82**, 122–137 (2017)
38. Antwi-Afari, M., Li, H., Edwards, D., Pärn, E., Seo, J., Wong, A.: Biomechanical analysis of risk factors for work-related musculoskeletal disorders during repetitive lifting task in construction workers. Autom. Constr. **83**, 41–47 (2017)
39. Lee, W., Lin, K.-Y., Seto, E., Migliaccio, G.C.: Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction. Autom. Constr. **83**, 341–353 (2017)
40. Han, S., Lee, S., Peña-Mora, F.: Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction. J. Comput. Civ. Eng. **28**, A4014005 (2013)
41. Ahmed, H., Tahir, M.: Improving the accuracy of human body orientation estimation with wearable IMU sensors. IEEE Trans. Instrum. Measur. **66**, 535–542 (2017)
42. Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., Havinga, P.: Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey. In: 2010 23rd International Conference on Architecture of Computing Systems (ARCS), pp. 1–10. VDE (2010)
43. Rawashdeh, S.A., Rafeldt, D.A., Uhl, T.L.: Wearable IMU for shoulder injury prevention in overhead sports. Sensors **16**, 1847 (2016)
44. Mannini, A., Sabatini, A.M.: Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors **10**, 1154–1175 (2010)
45. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) Pervasive 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24646-6_1
46. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3121–3124. IEEE (2010)

Visual Data and Predictive Analytics for Proactive Project Controls on Construction Sites

Jacob J. Lin(✉) and Mani Golparvar-Fard

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{jlin67,mgolpar}@illinois.edu

Abstract. This paper presents the theoretical foundation for a project controls system that improves understanding of how construction performance can be captured, communicated, and analyzed in form of a visual production system; predicts and effectively communicates the reliability of the weekly work plan and look-ahead schedules, supports root-cause assessment on plan failure at both project and task-levels; facilitates information flows; and decentralizes decision-making. Our model-driven system builds upon novel visual data analytics to map the current state of production in 4D (3D+time), compare to 4D BIM, and expose waste at both project and task-levels. Using predictive analytics and based on actual progress and productivity data, reliability in the future state of production is forecasted to highlight potential issues in a location-driven scheme and support collaborative decision making that eliminates root causes of waste. To evaluate the performance of our system, several case studies are conducted on real-world commercial building projects. It is shown that the developed system provides visual interfaces between people and information on and offsite, enables effective pull flows, decentralizes work tracking, facilitates in-process quality control and hand-overs among contractors, and most importantly transforms retroactive and task-driven workflows in contractor coordination meetings to proactive location-driven practices.

Keywords: Visual production management · Predictive data analytics
Lean construction

1 Introduction

Achieving smooth flow of production in construction requires team-based planning and systematic avoidance of waste through production control mechanisms [4, 7, 8, 11, 14, 24, 25, 36, 41]. Over the past decade, production control theories such as the Last Planner System [7] have emerged that stabilize workflows by shielding the direct work from upstream variation and uncertainty. While the benefits of these theories are well documented, yet their full potential across the life of a construction project is rarely achieved, and the root-causes for this are not fully understood.

A large body of recent studies suggest that successful implementation of control mechanisms, requires dedicated facilitators and engages practitioners in a relatively deep learning process [3, 13]. Sustaining this level of commitment for the duration of a project

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 412–430, 2018.
https://doi.org/10.1007/978-3-319-91635-4_21

is difficult, and in its absence, project teams revert back to traditional project control practices [9, 13, 24, 25, 35]. While these barriers are mostly attributed to the people and organizational processes involved in implementing lean principles, yet there is a growing recognition among researchers that the functional aspects of production control techniques need close re-examination to better *understand, predict,* and *analyze* reliability in performance, and *preserve effective and timely flow of information both to and from the workface* [4, 11, 13, 14, 24, 25, 36].

To address these knowledge gaps, *the overarching objective* of this research is to present the theoretical foundation for a project controls system that (a) improves understanding of how construction performance can be captured, communicated, and analyzed in form of a visual production system; (b) predicts and effectively communicates the reliability of the weekly work plan and look-ahead schedules, supports root-cause assessment on plan failure at both project and task-levels; (c) facilitates information flows; and (d) decentralizes decision-making. Specifically, this thesis presents a model-driven method built upon novel visual data sensing and analytics that maps the *current state* of production in 4D (3D+time) and exposes waste at both project and task-levels. Using predictive data analytics and based on actual progress and productivity data captured via visual data analytics and user input on worker-hr. data, reliability in the *future state* of production is forecasted to highlight potential issues in a location-driven scheme and support collaborative decision making that eliminates root causes of waste. The developed project controls system provides *visual interfaces between people and information* on and offsite and enables effective pull flow, decentralizes work tracking, and facilitates in-process quality control and hand-overs among contractors.

To ensure the implementation of this proactive visual production management system does not take away from actual productivity, the system extends the value of 4D Building Information Models (BIM) commonly used for constructability review [16] as a benchmark for performance. Likewise, it leverages images and videos, frequently collected [45] by project participants or professional services such as [15, 31, 38] via consumer-grade, time-lapse, smartphone cameras and Unmanned Aerial Vehicles (UAVs) to visually document actual performance on construction sites. The visual 4D Reality data is regularly captured and used together with 4D BIM to offer a unique opportunity to continuously compare and communicate Reality and Plan on construction sites and provide project teams with actionable project controls data analytics. Considering the current state of field reporting and information communication on construction sites, the application of the visual data for project controls has potential to significantly improve the transparency, accountability, and traceability of the communication between the field and the office.

In the following, an overview of the state of the construction industry is presented from a project controls perspective. Next, the specific practical problems that are the root-causes of the low productivity performance and the opportunity of using visual data as a source for capture, analytics and representation of Plan and As-built data on construction projects are discussed. In the next Chapter, state of the art in the literature that specifically focuses on addressing the project controls problems and the state of the art in visual data analytics for construction progress monitoring and controls are discussed in detail. Next, the overarching objective together with the specific research questions that this research addressed are discussed. The visual production management system and the underlying methods for developing 3D visual production maps, integration with 4D BIM, predictive data analytics, and representation of such data in a web-based client-server system architecture in used is discussed in detail. Several case studies and the finding of the research are presented as well.

1.1 The State of Productivity in the Construction Industry

Today, the construction industry is still plagued with inefficiencies, including cost overruns and delays in execution of projects. Productivity levels have remained flat for decades, particularly in comparison to other industries such as manufacturing, in which productivity doubled in the same period. According to a recent analysis by [12], 98% of mega projects – projects that the total construction cost is larger than 1 billion dollars - incur delays averaging 20 months and average cost overruns of 80%. This situation is not only limited to mega projects and also applied to all building projects. A recent report by Dodge Data and Analytics [37] also shows that more than 53% and 66% of common commercial and industrial building projects are behind schedule and over budget respectively. While best practices such as the Last Planner System and lean construction principles do improve schedule and cost performance on construction sites, still 22% and 13% of projects where best practices are implemented exhibit schedule delays and cost overruns respectively (Fig. 1).

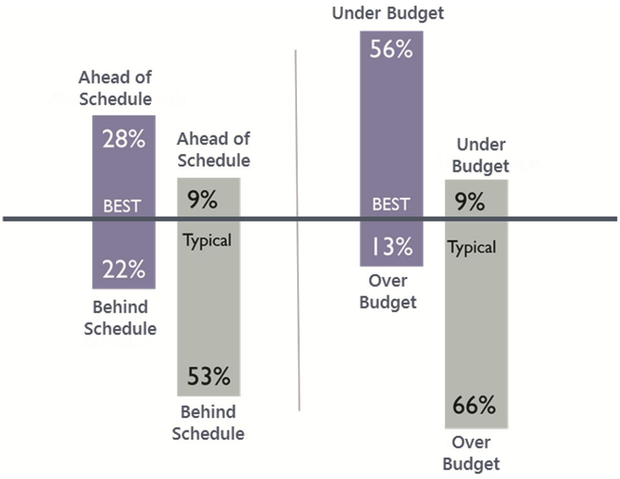


Fig. 1. Most typical commercial and industrial building projects complete behind schedule, over budget. Cost is improved more than schedule on best performing projects while best practices such as the Last Planner System and lean construction are introduced. Data from a report from ENR Dodge Data and Analytics (2016).

The same report [37] shows that among various project controls metrics, owners consider adherence to the project schedule is the number one issue that they face in the execution of their projects. There is a myriad of factors that have contributed to the lack of growth in construction productivity and the complexity of executing projects on time and on budget. A Careful examination of the most recent studies and reports including [12, 37] and internal anecdotal observations from more than 50 construction projects over the past 10 years that the University of Illinois at Urbana-Champaign has been involved in, has revealed a list of issues as key contributors and the root-causes of lack of productivity in the construction industry. The following, provides an overview of these specific problems:

1. **Inadequate communications.** Inconsistencies in reporting of project plans and actual work in place makes it difficult for subcontractors, contractors, and owners to maintain a common understanding of how projects are progressing at any given time. Quick and easy access to updated project information is key to resolving many of these factors that are leading to inconsistent reporting and ultimately cost and schedule difficulties. Construction monitoring improves efficiency because it ensures that construction plans are followed and that resources are fully utilized. However, current methods of construction monitoring can be expensive, time-consuming and subjective. In many instances, monitoring is performed infrequently and completely manually. As a result, companies are left with incomplete performance data that is unsuitable to use for accurate analytics.
2. **Flawed performance management.** Due to the lack of systematic and frequent communication and accountability in execution, the unresolved issues quickly stack up. As a result, teams are asked to present and explain progress achieved by their crews during weekly or bi-weekly contractor coordination meetings when significantly amount of working time is already lost to communicate and tap off issues.
3. **Poor short-term planning.** Construction firms are good at understanding and planning progress to be achieved in two to three month but rarely have an insight for next week or two. There are different levels of planning, from high-end preparation to day-by-day programs. If daily work does not go as planned, the scheduler should know about it but more often do not—so that they cannot update priorities in real time (Fig. 2).

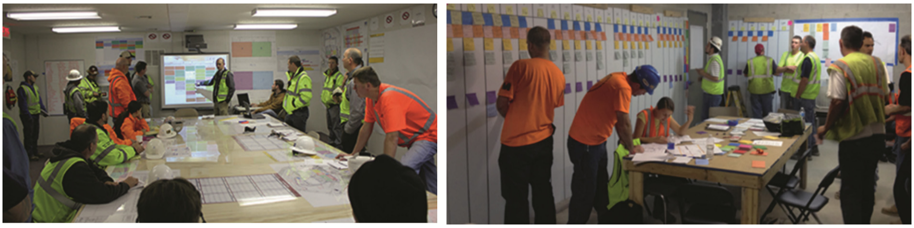


Fig. 2. Weekly contractor coordination and pull-planning meeting which range from high-end preparation to day-by-day schedule.

4. **Missed connections to actual progress.** The individuals involved in project planning or revising short-term and long-term plans are usually not working on construction sites or spent most of their time during the day sitting in jobsite trailers.

Naturally, the feedback and information that is expected from the site in terms of how much work is completed, and the actual productivity rates that could be achieved by teams is not provided to them, so there is no basis that could lead to improvement in project plans.

5. **Insufficient risk management.** Long-term risks get important consideration by management and project teams, however the kinds that crop up on the job not nearly as much. Particularly reliability and risk in short-term project plans are not systematically assessed.
6. **Poor decision-making.** Day-to-day planning and decision making is frequently inhibited due to poor communication surrounding daily work progress. Stagnant productivity levels can also be attributed to flawed performance management, poor short-term planning and decision-making, and insufficient short-term risk management. These problems associated with decision making are serious, systemic, and all too common in the construction industry [12]. Addressing these issues through timely and accurate information on project progress and performance can enable better communication and decision-making.

1.2 The Unprecedented Growth of Visual Data

The recent developments are the diffusion of consumer commodity devices with built-in cameras, such as smart phones, tablets, wearables, and camera-equipped unmanned ground/aerial vehicles (denoted as UAV). These camera-equipped platforms have led to an exponential growth in the volume of images and videos that are being recorded on construction sites on a daily basis [30]. To streamline the process of collecting images and videos, professional photography has also received significant attention. According to one of the popular construction documentation service providers, about 325,000 images are taken by professional photographers, 95,400 images by webcams, and 2000 images by construction project team members at a typical commercial building project (~750,000 sf). These are more than 400,000 images in total. The number can be much higher with the use of UAV for capturing images. The ever increasing volume of digital images provides an unprecedented opportunity to visually capture actual status of construction sites at a fraction of cost compared to other alternatives such as laser scanning.

In the meantime, Building Information Modeling (BIM) (i.e. 3D models enriched with performance information such as time, cost, safety, and productivity) has received a certain level of maturity. As BIM has advanced, enhanced 3D visualization with semantic building information at job sites improved communication and coordination. The applications include design development, construction coordination, and planning and the value-added by BIM are well documented. For example, Lu et al. (2014) report 6.92% cost saving by using BIM even after accounting for the added efforts during the design phase. Similarly, [43] report 25–30% productivity improvement through BIM-driven coordination and constructability reviews that identified most design conflicts before construction.

These emerging sources of visual data provide a unique opportunity to continuously reconstruct and visualizes Reality directly within 4D BIM, measures progress and productivity, and analyzes risk for delay. By putting schedule tasks and project performance data in a visual context for the entire team and mapping Reality to Plan, a visual production management platform provides transparency in project execution and helps project team better plan, coordinate, and communicate.

In the next section, the state of the art in the literature that specifically focuses on addressing the project controls problems and the state of the art in visual data analytics for construction progress monitoring and controls are discussed in detail. Next, the overarching objective together with the specific research questions that this research addressed are discussed. The visual production management system and the underlying methods for developing 3D visual production maps, integration with 4D BIM, predictive data analytics, and representation of such data in a web-based client-server system architecture in used is discussed in detail. Several case studies and the finding of the research are presented as well.

2 Related Work

2.1 Work Status Communication

Communicating "who does what work in what location" is key in tracking state of work-in-progress, resource allocation, task readiness and revising project schedules [13, 34]. Achieving this level of communication requires significant upfront efforts to assemble the Work Breakdown Structure in BIM for work tracking [18]. Communicating and documenting what is done and what should and can be done still relies on going back and forth between the field office and site to access the paper-based documents [5, 32, 33], or searching on smartphones which needs manual work to generate feasible view for each task [6]. As a result, systematically connecting the input (worker-hours, equipment and material) to the output (progress) is still very time-consuming and hard to achieve.

2.2 Progress Monitoring

To improve work status communication, research has focused on automating the process of progress monitoring through comparison between 4D BIM and time-lapse videos [1, 2, 19, 45], or 3D image-based or laser scanning point clouds [10, 19, 20, 22, 42, 44]. These methods utilize physical appearance of building elements captured by camera or laser scanner and relate that to BIM to detect progress deviations. As a result, they rely on high accuracy of geometry [19] or formalized knowledge of sequencing and reasoning mechanism via appearance-based recognition [28, 29]. At best, these methods still rely on retroactive Earned Value Analysis and as such do not proactive communicate potential performance problem. Therefore, work status is not communicated in time to prepare superintendents to mobilize their crew into new work locations. The lack of a systematic reporting on task status, methods, resources delays, inspections, safety check-ins and crew mobilization ultimately leads to waste.

2.3 Work Tracking

Research has utilized 4D BIM for pull planning and decentralizing documentation of work status [18, 24], however the process of reporting progress has remained a manual and time-consuming process. In addition, assigning namespaces [18] for tracking work in progress does not easily scale to cases where work breakdown structure and model breakdown structures are inconsistent.

2.4 Predictive Data Analytics

Today, PPC (Percent Planned Complete) and Earned Value Analysis Schedule and Cost Performance Indexes (SPI/CPI) are widely used for project controls purposes. However, these metrics are retrospective and only enable learning from the past mistakes [24, 34]. Recent research proposes proactively measuring reliability of the weekly work plan and look-ahead plan through implementing *time-dependent maturity index* [39], *Task Anticipated* and *Task Made Ready* metrics [26, 27]. These metrics require real-time feedback from the site on actual progress and productivity which is rare.

2.5 Visual Data Applications

Generating large panoramic images of the site and superimposing these large-scale high resolution images over existing maps – While these images provide excellent visuals to ongoing operation (Fig. 3), they lack 3D information to assist with area-based and volumetric-based measurements necessary for progress monitoring. Also none of the current commercially available platforms provide a mechanism to communicate *who* is working on *which* tasks at *what* location and they mainly deliver high quality maps of construction sites.

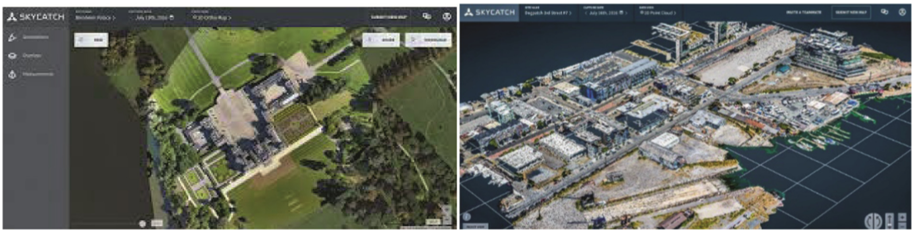


Fig. 3. Skycatch drone based visual data management platform – high-level top-down images are used to produce large-scale high resolution orthophotos and overlay them over existing maps. These images are also used to generate point cloud models.

To address these limitations, a new system proactive project controls is presented that enables lean pull strategies through decentralizing the work tracking and visualizing work status, availability of resources, and readiness of tasks in the look-ahead schedule. The details of this system are provided in the following section.

3 Visual Production Management System

By leveraging the unprecedented visual data collected on construction site through application of smartphones, fixed cameras and drones, this section introduces the underlying visual production management system in the envisioned proactive project controls system. Specifically, a new visual asset management system is introduced which takes in images captured at different times and locations to automatically generate 4D as-built point clouds and also localize unordered images in the same environment. By linking look ahead schedules and BIM via Work Breakdown Structure locations, location-based 4D Models are created and merged with the visual asset management platform to benchmark and monitor "who does what work in what location". By accessing the resulting visual production models through smartphones, worker-hours are documented per task per organization and per location (Fig. 4). The reliability of the look ahead schedule is measured based on actual production and productivity rates and top locations at risk for potential delays are highlighted on a weekly basis. A risk-driven workflow is also introduced to helps project team tap off potential delays in weekly coordination meetings. The following presents the details of each method used in the visual production management system.

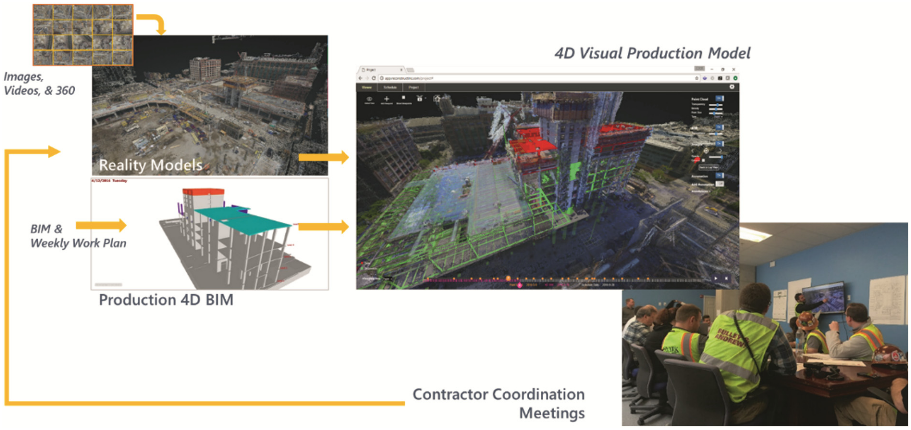


Fig. 4. The Visual Production Management System is generated by continuously reconstructs and visualizes 4D Reality from visual data collected through commodity smart phones, fixed camera, UAVs, videos and 360 camera; and integrates with production 4D BIM and used during coordination meeting for planning, communication and coordination.

To ensure the implementation of this proactive visual production management system does not take away from actual productivity, the system extends the value of 4D Building Information Models (BIM) commonly used for constructability review [16] as a benchmark for performance. Likewise, it leverages images and videos, frequently collected [45] by project participants or professional services such as [15, 31, 38] via consumer-grade, time-lapse, smartphone cameras and Unmanned Aerial Vehicles (UAVs) to visually document actual performance on construction sites. The visual 4D

Reality data is regularly captured and used together with 4D BIM to offer a unique opportunity to continuously compare and communicate Reality and Plan on construction sites and provide project teams with actionable project controls data analytics. Considering the current state of field reporting and information communication on construction sites, the application of the visual data for project controls has potential to significantly improve the transparency, accountability, and traceability of the communication between the field and the office.

Given a collection of visual data captured by drone, fixed camera, digital camera or commodity smart phone, the first goal is to automatically reconstruct the 3D As-Built model to map the current state of construction. The State-of-the-art is to first reconstruct the construction site using image-based 3D reconstruction and then registering the BIM model with the point cloud to localize all the images with BIM. An image-based geometry reconstruction pipeline which consists of Graphic Processing Unit (GPU)-based Structure-from-Motion (SfM) [23], Multi-View Stereo (MVS) and Surface Reconstruction [17] is introduced to automatically generate a dense 3D point cloud of the current state of construction. 4D Point clouds are generated from images collected from different time and aligned with each other using distinguished visual feature that remain the same across different time of the construction site. The BIM model is registered with the point clouds using markers placed on the construction site with surveying points benchmarked with the coordination in the BIM model, or directly using features recognized on the point cloud with BIM. 4D BIM is also generated to allowing the last planners dynamically and visually commit to task and location. A location-based 4D BIM is introduced in this system. In the following sections, each step will be described in detail.

The construction site is changing tremendously over time and the collected image content during the time period will also be different. However, many elements on the site remain unchanged such as finished elements and existing structure around the site. Leveraging these unchanged elements, the SfM process is capable to generate automatically aligned point cloud. In other words, every time a new set of images are uploaded in the system, these images are matched with previously generated point cloud models and are localized in the same coordinate system. The newly reconstructed point cloud, as a result, will also be automatically aligned with the previously generate model. The detail process is described in Fig. 5.

If the SfM process fail for newer point cloud, it can still be registered using the absolute transformation [21] to the on scale point cloud coordinates. Once these models are generated, they are mapped into a timeline (typically linked to the project schedule) to reflect the timing of their capture. Figure 6 provides an example of these point cloud models produced and mapped at McCormick Place construction site in Chicago, IL. As can be seen from the figures, images captured during different days and under various lighting conditions can all be used to produce 3D point cloud models and bring all images and point clouds into alignment. The point cloud generation are fully automated. The point cloud registration to BIM can be performed automatically if survey benchmarks and survey marker or fiduciary marker are placed onsite, the registration process can also be performed in the system manually. The 4D BIM generation are semi-automated depends if the codes between the model and schedule are created.

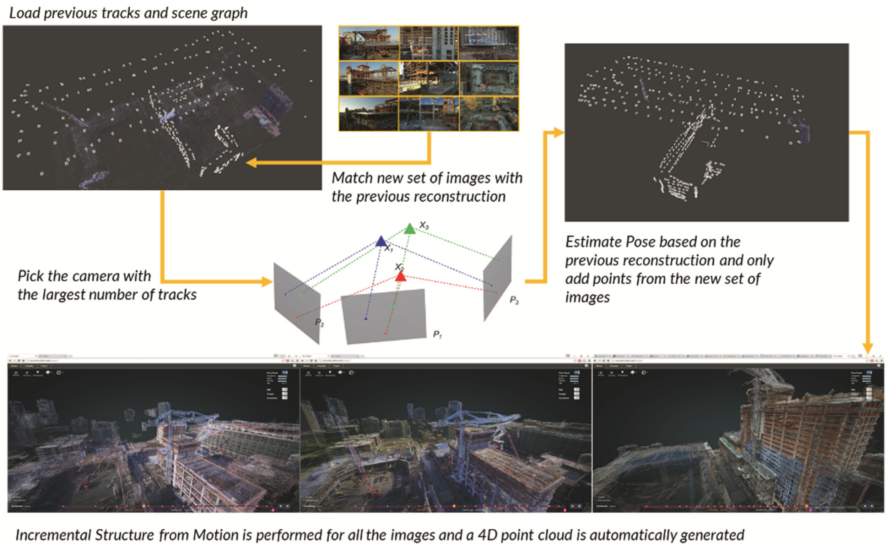

Load previous tracks and scene graph
Match new set of images with the previous reconstruction
Estimate Pose based on the previous reconstruction and only add points from the new set of images
Pick the camera with the largest number of tracks
X_3
X_1
X_2
P_2
P_1
P_3
Incremental Structure from Motion is performed for all the images and a 4D point cloud is automatically generated

Fig. 5. The 4D point cloud is automatically generated by utilizing the previous reconstructions. The previous tracks and scene graph are first loaded, and features from the new images are then match with the previous reconstruction. Based on the matching, we pick the camera that has the largest number of tracks, and register the camera with the previous point cloud. We estimate the pose of the camera and triangulate all the points that can be observed by more than two cameras. Only points from the new set of images will be added. This way the new reconstruction will be in the same coordinates as the previous reconstruction, and a 4D point cloud is automatically generated.

Inspired by [40] and instead of simply measuring and tracking PPC, we introduce metrics for measuring reliability of current weekly work plans and two-week look ahead schedules. The key idea was that PPC is a retroactive performance metric and only allows learning from past performances. Nevertheless in construction projects, tasks do not repeat to enable learning from the past. Here we use Percent Complete (PC) for tasks and task constraints. If the readiness level associated with upcoming tasks are measured, teams will have a better understanding of which tasks can start on time and which ones require revision in upcoming coordination meetings. The metrics, Readiness Index (*RI*) and Readiness Reliability (*RR*) for each ongoing and/or upcoming task i simply measures the status of all task predecessors and their procurement and logistical constraints j to provide construction teams with a better understanding of when each team can mobilize their crews to initiate certain tasks. When these tasks are not ready for a scheduled start time, the teams will be able to pull those tasks that are ready and in turn stabilize workflows. The resulted number of the metrics are shown in the web-based platform via color-coded model to communicate the readiness of the tasks. Equations 1 to 3 represents how these metrics can be measured:

$$RI_i = \sum\nolimits_j \varphi_j \,.\, PC_j; \; \sum\nolimits_j \varphi_j = 1 \tag{1}$$

$$\mathrm{RR_i} = RI_i + \sum\nolimits_j \varphi_j . (1 - PC_j) . \Omega_{mj} \tag{2}$$

$$\Omega_{\mathrm{mj}} = \mathrm{P}(t \le T) = 1 - e^{-\lambda t} \tag{3}$$

where Ω_{mj} presents the probability of contractor (m) finishing the remaining task, and PC_j is the percent complete on task j. The weight function φ is based on the experience of the project engineers to decide the importance of the task. However, it will be divided to equally to all the predecessor tasks and task constraints if it is not assigned. Here we use Poisson distribution to estimate the expected duration via the prior task executed by the same contractors in Eq. 3. Parameter λ will be calibrated via expected durations based on productivity and a normalized penalty term based on observed delays in all prior tasks executed by the contractor or supplier on a project. The productivity are collected through the production level 4D BIM and the progress report from the 4D point cloud and images. The metrics are shown in the web-based viewer and reports to proactively communicate the progress and readiness of schedule (Fig. 9).

4 Validation

To validate the visual production management system, we partnered with nationally recognized construction companies to implement our system in several projects. Qualitative evaluation has been made through observation and interviews. We have observed the user utilizing the system during coordination meetings and field operation. One representative project with preliminary results is discussed in depth in the following.

This pilot project was conducted with an ENR US Top 20 General Contractor on a combined cost of $500 million hotel and event center construction. A forty-story hotel tower and a 300,000 square feet event center project utilized the visual production management system to improve the overall productivity of the project. Over a span of 4 months, drone flight was operated 27 times on a weekly basis to collect exterior visual data. Interior visual data is also collected over site walks performed by project engineers over multiple weeks. A workflow from visual data capture and process, predictive data analytics generation, at-risk report review during project management meeting, field engineers committing to task, and real-time 4D discussion during coordination meetings has been established and implemented. We specifically focused on this workflow that identifies, characterizes, and communicates actual and potential performance problems. Every week the project teams were provided with visual production reports which highlighted the top 10 at-risk locations (locations where tasks will likely suffer from potential performance problems), together with all ongoing tasks and their percent complete rates, a list of upcoming tasks with respect to the schedule in that location, their readiness levels (as well as their logistical and contractual constraints) and the location's stability index. A sequence model (Fig. 7) is generated to show the work flow of the original coordination meetings and the mapping of the system features that can facilitate the meeting process based on roles. The general contractor moderates the meeting and go through the tasks based on subcontractors or locations. They start with discussing the upcoming activities and constraints to adjust the duration and finally commit to the task.

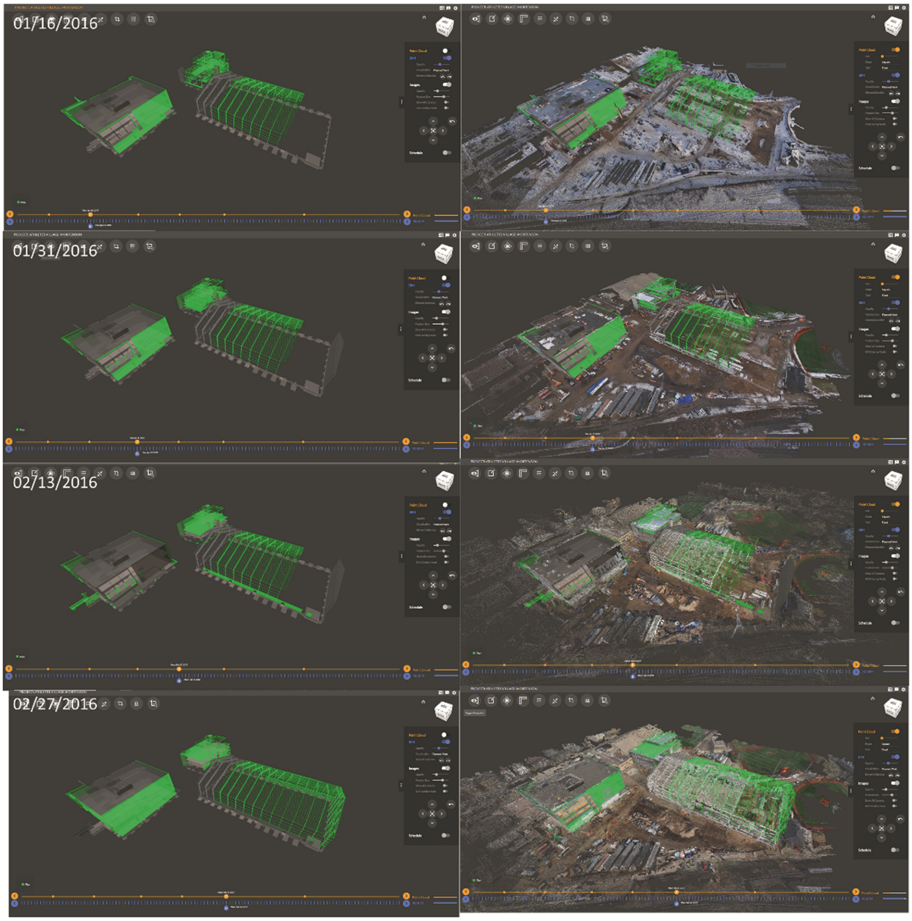
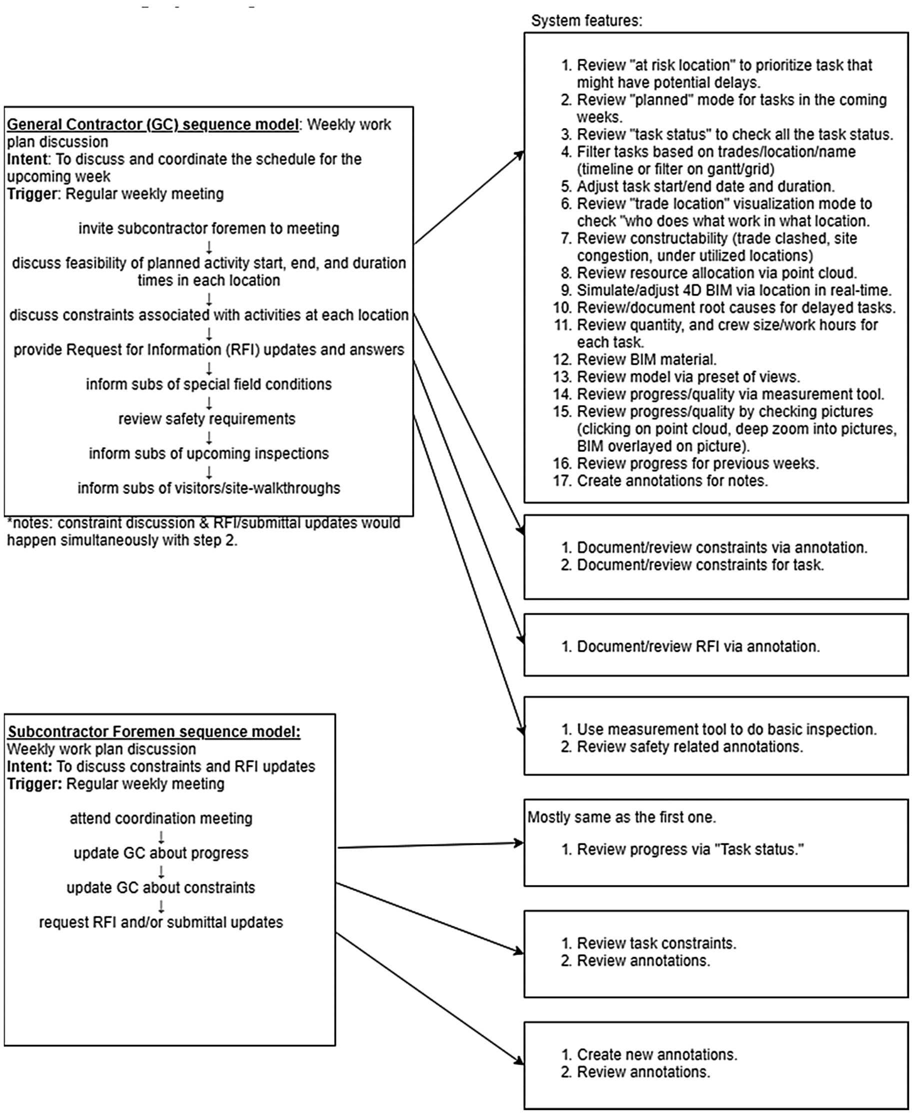


Fig. 6. In the left side shows a 4D BIM which is updated directly during the weekly coordination meetings to simulate the work for the coming weeks; in the right shows the visual production model which integrated the 4D BIM model with the 4D point cloud.

The sequence model is generated based on the interviews with superintendents and project engineers.

After the system was introduced to the system, the top 10 at-risk location are first used during the project management and scheduling meeting to prioritize discussion of possible project action that can improve the progress. During the field supervision meeting with the subcontractor, schedule are already revised based on the at risk location so that when subcontractor commit to tasks, the duration of the task are more close to the reality. When communicating the tasks to the subcontractors, “who does what work in what location” is completely communicate through the system (Fig. 8), and subcontractor can directly make comment and feedback to the progress, which transform pushing to pulling.



General Contractor (GC) sequence model: Weekly work plan discussion
Intent: To discuss and coordinate the schedule for the upcoming week
Trigger: Regular weekly meeting

invite subcontractor foremen to meeting
↓
discuss feasibility of planned activity start, end, and duration times in each location
↓
discuss constraints associated with activities at each location
↓
provide Request for Information (RFI) updates and answers
↓
inform subs of special field conditions
↓
review safety requirements
↓
inform subs of upcoming inspections
↓
inform subs of visitors/site-walkthroughs

*notes: constraint discussion & RFI/submittal updates would happen simultaneously with step 2.

System features:

1. Review "at risk location" to prioritize task that might have potential delays.
2. Review "planned" mode for tasks in the coming weeks.
3. Review "task status" to check all the task status.
4. Filter tasks based on trades/location/name (timeline or filter on gantt/grid)
5. Adjust task start/end date and duration.
6. Review "trade location" visualization mode to check "who does what work in what location.
7. Review constructability (trade clashed, site congestion, under utilized locations)
8. Review resource allocation via point cloud.
9. Simulate/adjust 4D BIM via location in real-time.
10. Review/document root causes for delayed tasks.
11. Review quantity, and crew size/work hours for each task.
12. Review BIM material.
13. Review model via preset of views.
14. Review progress/quality via measurement tool.
15. Review progress/quality by checking pictures (clicking on point cloud, deep zoom into pictures, BIM overlayed on picture).
16. Review progress for previous weeks.
17. Create annotations for notes.

1. Document/review constraints via annotation.
2. Document/review constraints for task.

1. Document/review RFI via annotation.

1. Use measurement tool to do basic inspection.
2. Review safety related annotations.

Subcontractor Foremen sequence model:
Weekly work plan discussion
Intent: To discuss constraints and RFI updates
Trigger: Regular weekly meeting

attend coordination meeting
↓
update GC about progress
↓
update GC about constraints
↓
request RFI and/or submittal updates

Mostly same as the first one.

1. Review progress via "Task status."

1. Review task constraints.
2. Review annotations.

1. Create new annotations.
2. Review annotations.

Fig. 7. A sequence model of the coordination meeting with the mapping of the system features. It clearly shows how the system facilitates different steps in the coordination meeting that identifies, characterizes, and communicates actual and potential performance problems.

The project was delayed by two months before the team start to use the visual production management system. We used Plan Percent Complete (PPC), tasks delayed to measure how the system has improved the communication of tasks after the system is introduced. PPC has increased in a favorable trend and almost the PPC of all the weeks after the system introduced is above the national average. Specifically, the PPC increased 24% from a 46% baseline in 6 weeks and remains above the national average. The tasks delayed decreased from 6 to 2 tasks per week, and the average repeated tasks are reduced from 12 to 4 tasks per week which indicates the subcontractor becomes more committed to finish the tasks according to the plan.

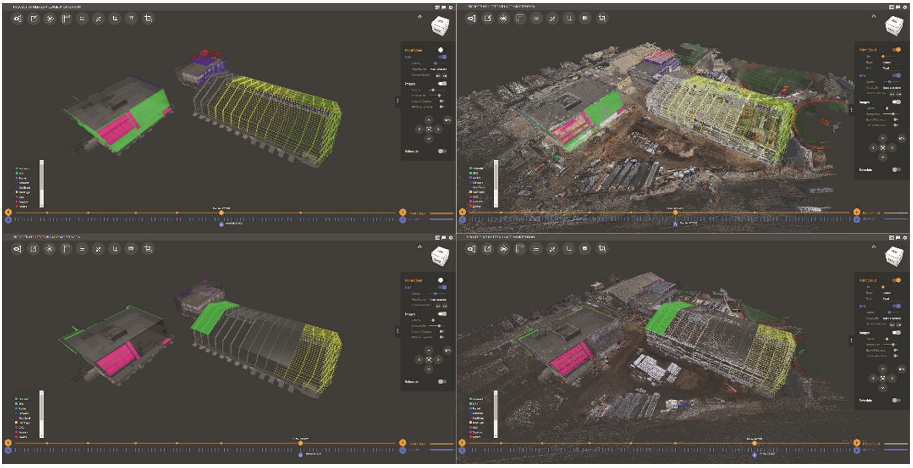


Fig. 8. The first column shows progress via 4D point cloud and BIM; the second column shows the 4D BIM with sub-contractor responsible tasks color-coded to communicate who does what work in what location.

Overall, the system offers three main improvements to the current workflow. First, it provides predictable and reliable plans through indicating the top at-risk locations. Highlighted locations and workflows are calculated by predictive analytics and can be prioritized during the coordination meetings. This prevents changes in work plan and improves the schedule reliability (Fig. 9). Second, it addresses delays caused by occupied work area and incomplete previous handoff through visually communicate “Who is doing What Work in What Location.” Subcontractors can easily coordinate and communicate the work plan for each day with others visually and report the progress back to the system (Fig. 10). This with visual data delivers real-time knowledge about task completion, availability of work space and task backlogs to commit. Third, it avoids under-estimated effort for tasks during coordination and planning through tracking production and productivity rates per work package or trade. Gauging productivity throughout the projects also provides the opportunity to analyze how resources are spent to increase the performance. These daily reports integrated with visual data are then transformed into actionable data to measure progress and reflect on planning (Figs. 11 and 12).

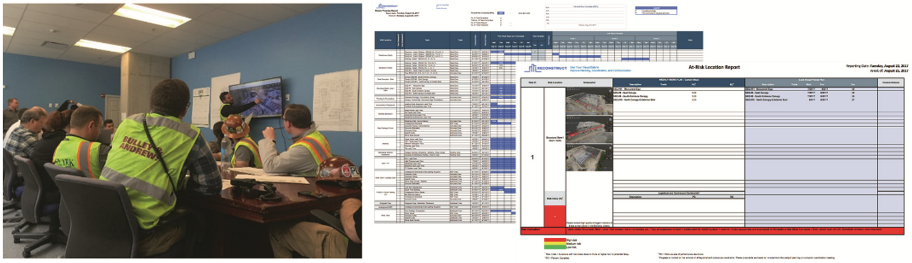


Fig. 9. Highlighted locations and workflows are calculated by predictive analytics and can be prioritized during the coordination meetings. This prevents changes in work plan and improves the schedule reliability.

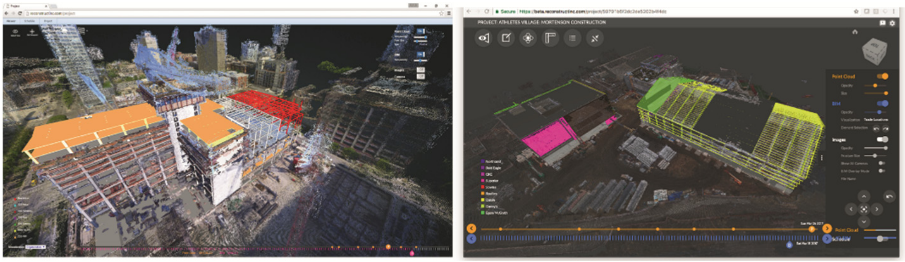


Fig. 10. Occupied work area and incomplete previous handoff are avoided through visually communicate "Who is doing What Work in What Location.

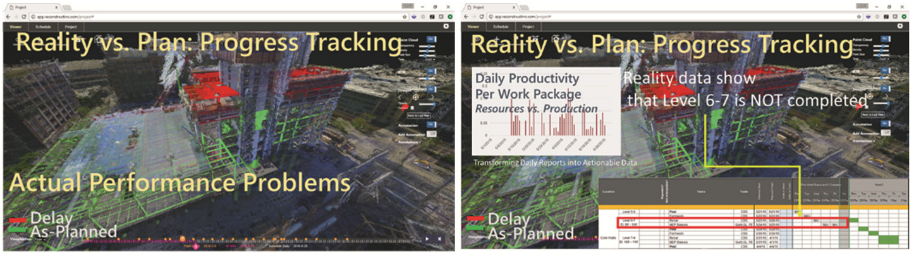

Reality vs. Plan: Progress Tracking
Actual Performance Problems
Delay
As-Planned
Reality vs. Plan: Progress Tracking
Daily Productivity
Per Work Package
Resources vs. Production
Reality data show
that Level 6-7 is NOT completed
Transforming Daily Reports into Actionable Data
Delay
As-Planned

Fig. 11. Production and productivity rates are tracked per work package or trade to be transformed into actionable data.

Fig. 12. The system has been used on different construction site during the coordination meetings, it has been proved that it can efficiently enhance accountability and traceability, and predictive analytics improve reliability in short-term planning.

5 Conclusion

An effective production management system to systematically plan and prevent waste is imperative to keep a smooth flow of the production in construction. Nonetheless, current production control theories such as Last Planner System has rarely achieved the full potential of the benefits it could bring. To address the challenges, this paper aims to developed a production management system that (a) improves understanding of how construction performance can be captured, communicated, and analyzed in form of a visual production system; (b) predicts and effectively communicates the reliability of the weekly work plan and look-ahead schedules, supports root-cause assessment on plan failure at both project and task-levels via the new metrics; (c) facilitates information flows; and (d) decentralizes decision-making. Preliminary results shows that the implementation of the system can improve transparency by visualizing and communicating both actual and potential problems as well as production and productivity rates; accountability by visualizing and communicating who is expected to do what work in what location; and traceability by providing a baseline for tracking performance on construction sites. The reliability of the weekly work plan and look-ahead schedules are also effectively communicates through color-coded model in the visual production model.

The visual production management system leverages visual data to improve planning, coordination and communication. It provides transparent process view to eliminate potential problems, more predictable plans (PPC increased by 24% over 46% baseline in 6 weeks) and high stability of personnel planning and logistics. It can also be further utilized for streamlined billing by verification of work, enhanced quality control through comparison of as-built and safer operations through knowledge of resources and safety hazards. Future works contains more case studies specifically on interior works. While the system can support for interior progress tracking as long as point clouds or images are provided, and preliminary research has also been done for interior spaces, interior image-based 3D reconstruction remains more challenging in terms of the practicality on construction site.

Acknowledgement. This material is in part based upon work supported by the National Science Foundation Grant #1446765. The support and help of our industrial collaborators in collecting data and implementing the work tracking system is greatly appreciated. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not reflect the views of the NSF, or the industrial partner mentioned above.

References

1. Abeid, J., Allouche, E., Arditi, D., Hayman, M.: PHOTO-NET II: a computer-based monitoring system applied to project management. Autom. Constr. **12**, 603–616 (2003). https://doi.org/10.1016/S0926-5805(03)00042-6
2. Abeid, J., Arditi, D.: Linking time-lapse digital photography and dynamic scheduling of construction operations. J. Comput. Civ. Eng. **16**, 269–279 (2002). https://doi.org/10.1061/(asce)0887-3801(2002)16:4(269)
3. Alarcón, L.F., Diethelm, S., Rojo, O., Calderón, R.: Evaluando los impactos de la implementación de lean construction. Rev. Ing. Constr. **23**, 26–33 (2008)
4. Aritua, B., Smith, N.J., Bower, D.: Construction client multi-projects–a complex adaptive systems perspective. Int. J. Proj. Manag. **27**, 72–79 (2009)
5. Bae, H., Golparvar-Fard, M., White, J.: High-precision vision-based mobile augmented reality system for context-aware architectural, engineering, construction and facility management (AEC/FM) applications. Vis. Eng. **1**, 3 (2013)
6. Bae, H., Golparvar-Fard, M., White, J.: High-precision vision-based mobile augmented reality system for context-aware architectural, engineering, construction and facility management (AEC/FM) applications. Vis. Eng. **1**, 1–13 (2013)
7. Ballard, G.: The last planner system of production control. The University of Birmingham (2000)
8. Ballard, G., Howell, G.: Implementing lean construction: stabilizing work flow. Lean Constr. (1994)
9. Bortolazza, R.C., Costa, D.B., Formoso, C.T., et al.: A quantitative analysis of the implementation of the last planner system in Brazil. In: 13th Conference of the International Group for Lean Construction, pp. 413–420. International Group on Lean Construction (2005)
10. Bosché, F., Guillemet, A., Turkan, Y., Haas, C., Haas, R.: Tracking the built status of MEP works: assessing the value of a Scan-vs-BIM System. J. Comput. Civ. Eng. (2013)
11. Brodetskaia, I., Sacks, R., Shapira, A.: Stabilizing production flow of interior and finishing works with reentrant flow in building construction. J. Constr. Eng. Manag. **139**, 665–674 (2013). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000595
12. Changali, S., Azam, M., van Nieuwland, M.: The construction productivity imperative (2015)
13. Dave, B., Hämäläinen, J.-P., Koskela, L.: Exploring the recurrent problems in the last planner implementation on construction projects, pp. 1–10 (2015)

14. Dave, B., Kubler, S., Främling, K., Koskela, L.: Addressing information flow in lean production management and control in construction. In: Proceedings of the 22nd International Group for Lean Construction, vol. 2, pp. 581–592 (2014)
15. Earthcam: EarthCam - Webcam Network (2015)
16. Eastman, C., Teicholz, P., Sacks, R., Liston, K.: BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors. John Wiley & Sons, Hoboken (2011)
17. Fuhrmann, S., Langguth, F., Goesele, M.: MVE: a multi-view reconstruction environment. In: Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage. Eurographics Association, Aire-la-Ville, Switzerland, pp. 11–18 (2014)
18. Garcia-Lopez, N.P., Fischer, M.: A system to track work progress at construction jobsites. In: The Industrial and Systems Engineering Research Conference (2014)
19. Golparvar-Fard, M., Peña-Mora, F., Arboleda, C.A., Lee, S.: Visualization of construction progress monitoring with 4D simulation model overlaid on time-lapsed photographs. J. Comput. Civ. Eng. **23**, 391–404 (2009)
20. Golparvar-Fard, M., Pena-Mora, F., Savarese, S.: D4AR-A 4-Dimensional augmented reality model for automating construction progress data collection, processing and communication. J. ITCON – Spec. Issue Next Gener. Constr. IT Technol. Foresight Futur. Stud. Roadmapping Scenar. Plan **14**, 129–153 (2009)
21. Golparvar-Fard, M., Peña-Mora, S., Savarese, S.: Automated model-based progress monitoring using unordered daily construction photographs and IFC as-planned models. ASCE J. Comput. Civ. Eng. (2012)
22. Golparvar-Fard, M., Peña-Mora, S., Savarese, S.: Automated model-based progress monitoring using unordered daily construction photographs and IFC as-planned models. ASCE J. Comput. Civ. Eng. **147** (2012). https://doi.org/10.1061/(asce)cp.1943-5487.0000205
23. Golparvar-Fard, M., Peña-Mora, F., Savarese, S.: Automated progress monitoring using unordered daily construction photographs and IFC-Based building information models. J. Comput. Civ. Eng. **147** (2012)
24. Gurevich, U., Sacks, R.: Examination of the effects of a KanBIM production control system on subcontractors' task selections in interior works. Autom. Constr. **37**, 81–87 (2014). https://doi.org/10.1016/j.autcon.2013.10.003
25. Hamzeh, F., Ballard, G., Tommelein, I.: Rethinking lookahead planning to optimize construction workflow. Lean Constr. J., 15–34 (2012)
26. Hamzeh, F., Bergstrom, E.: The lean transformation: a framework for successful implementation of the last PlannerTM system in construction. In: Proceedings of the 46th Annual International Conference of the Associated Schools of Construction (2010)
27. Hamzeh, F.R., Saab, I., Tommelein, I.D., Ballard, G.: Understanding the role of "tasks anticipated" in lookahead planning through simulation. Autom. Constr. **49**, 18–26 (2015). https://doi.org/10.1016/j.autcon.2014.09.005
28. Han, K.K., Cline, D., Golparvar-Fard, M.: Formalized knowledge of construction sequencing for visual monitoring of work-in-progress via incomplete point clouds and low-LoD 4D BIMs. Adv. Eng. Inform. **29**, 889–901 (2015). https://doi.org/10.1016/j.aei.2015.10.006
29. Han, K.K., Golparvar-Fard, M.: Appearance-based material classification for monitoring of operation-level construction progress using 4D BIM and site photologs. Autom. Constr. **53**, 44–57 (2015)
30. Han, K.K., Golparvar-Fard, M.: Potential of big visual data and building information modeling for construction performance analytics: an exploratory study. Autom. Constr. **73**, 184–198 (2017). https://doi.org/10.1016/j.autcon.2016.11.004

31. JobSiteVisitor: JobsiteVisitor.com (2015)
32. Kamat, V.R., Akula, M.: Integration of global positioning system and inertial navigation for ubiquitous context-aware engineering applications. In: Proceedings of the National Science Foundation Grantee Conference, Atlanta, GA, pp. 1–10
33. Kamat, V.R., Martinez, J.C., Fischer, M., Golparvar-Fard, M., Peña-Mora, F., Savarese, S.: Research in visualization techniques for field construction. J. Constr. Eng. Manag. **137**, 853–862 (2010)
34. Koskela, L., Howell, G.: The underlying theory of project management is obsolete. IEEE Eng. Manag. Rev. **36**, 22–34 (2008). https://doi.org/10.1109/EMR.2008.4534317
35. Leigard, A., Pesonen, S.: Defining the path - a case study of large scale implementation of last planner. In: Proceedings of the 18th Annual Conference of the International Group for Lean Construction, vol. 1, pp. 1–10 (2010)
36. Lindhard, S., Wandahl, S.: Improving onsite scheduling: looking into the limits of last planner system. Built Hum. Environ. Rev. **6**, 46–60 (2013)
37. Mace, B., Balfour, B., Jones, S.: How satisfied, really satisfied, are Owners? (2016)
38. MultiVista: Multivista Construction Documentation Photos, Videos & WebCams (2015)
39. Sacks, R., Barak, R., Belaciano, B., Gurevich, U., Pikas, E.: Kanbim workflow management system: prototype implementation and field testing. Lean Constr. J. **9**, 19–34 (2013)
40. Sacks, R., Koskela, L., Dave, B.A., Owen, R.: Interaction of lean and building information modeling in construction. J. Constr. Eng. Manag. **136**, 968–980 (2010). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000203
41. Sacks, R., Radosavljevic, M., Barak, R.: Requirements for building information modeling based lean production management systems for construction. Autom. Constr. **19**, 641–655 (2010). https://doi.org/10.1016/j.autcon.2010.02.010
42. Son, H., Bosché, F., Kim, C.: As-built data acquisition and its use in production monitoring and automated layout of civil infrastructure: a survey. Adv. Eng. Inform. **29**, 172–183 (2015). https://doi.org/10.1016/j.aei.2015.01.009
43. Staub-French, S., Khanzode, A.: 3D and 4D modeling for design and construction coordination: Issues and lessons learned. ITcon **12**, 381–407 (2007). http://www.itcon.org/2007/26
44. Turkan, Y., Bosche, F., Haas, C., Haas, R.: Automated progress tracking using 4D schedule and 3D sensing technologies. Autom. Constr. **22**, 414–421 (2012)
45. Yang, J., Park, M.-W., Vela, P.A., Golparvar-Fard, M.: Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future. Adv. Eng. Inform. **29**, 211–224 (2015)

Quantification of Energy Consumption and Carbon Dioxide Emissions During Excavator Operations

Hassanean S. H. Jassim[1,2](✉), Weizhuo Lu[1], and Thomas Olofsson[1]

[1] Lulea University of Technology, 97187 Lulea, Sweden
hassanean.jassim@ltu.se
[2] Babylon University, 51002 Babylon, Iraq

Abstract. A number of studies have assessed the energy consumed and carbon dioxide emitted by construction machinery during earthwork operations. However, little attention has been paid to predicting these variables during planning phases of such operations, which could help efforts to identify the best options for minimizing environmental impacts. Excavators are widely used in earthwork operations and consume considerable amounts of fuel, thereby generating large quantities of carbon dioxide. Therefore, rigorous evaluation of the energy consumption and emissions of different excavators during planning stages of project, based on characteristics of the excavators and projects, would facilitate selection of optimal excavators for specific projects, thereby reducing associated environmental impacts. Here we describe use of artificial neural networks (ANNs), developed using data from Caterpillar's handbook, to model the energy consumption and CO_2 emissions of different excavators per unit volume of earth handled. We also report a sensitivity analysis conducted to determine effects of key parameters (utilization rate, digging depth, cycle time, bucket payload, horsepower, load factor, and hauler capacity) on excavators' energy consumption and CO_2 emissions. Our analysis shows that environmental impacts of excavators can be most significantly reduced by improving their utilization rates and/or cycle times, and reducing their engine load factor. We believe our ANN models can potentially improve estimates of energy consumption and CO_2 emissions by excavators. Their use in planning stages of earthworks projects could help planners make informed decisions about optimal excavator(s) to use, and contractors to evaluate environmental impacts of their activities. Finally, we describe a case study, based on a road construction project in Sweden, in which we use empirical data on the quantities and nature of the materials to be excavated, to estimate the environmental impact of using different excavators for the project.

Keywords: Energy use and CO_2 emission
Simulation and ANN predicting model · Sensitivity analysis

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 431–453, 2018.
https://doi.org/10.1007/978-3-319-91635-4_22

1 Introduction

Despite rapid developments in industrial technology and machine manufacturing in recent decades, construction equipment is still a major contributor to environmental impacts due to the large fleet of old machinery that will remain in service for several years [1]. Non-road diesel engines and equipment reportedly consume approximately 70% as much diesel as road transport vehicles in Australia [2] (for example), and emissions of CO_2 (a major 'greenhouse gas', reportedly accounting for about 60% of total anthropogenic global warming) are directly related to the amount of fuel consumed [3, 4]. Moreover, numerous studies have shown that the construction industry is responsible for major proportions of total fuel consumption and emissions, and that excavators make substantial contributions to the environmental impact of the construction sector [5]. For example, an extensive study on construction equipment in the USA found that excavators accounted for 15% of the total energy consumption and CO_2 emissions from construction equipment and machinery [6]. Hence, total emissions from construction machinery and equipment could be reportedly reduced by up to 10% through careful selection of machinery [7] and optimizing its usage time. Therefore, developed countries such as The Netherlands [8] and Sweden [9] are making efforts to reduce the environmental impact of road and infrastructure projects. For example, the Swedish Transport Administration (STA) has recently stated that efforts to mitigate energy use and (CO_2) emissions from construction infrastructure projects will be prioritized [9], and incentives will be provided for contractors to mitigate emissions during planning and construction.

Clearly, rigorous evaluation of the energy used and emissions (CO_2) from excavators could significantly help efforts to mitigate the environmental impact of construction equipment. Thus, a number of studies have considered energy consumption and emissions of construction equipment, using various approaches or models [10–16]. However, they have not provided robust techniques for predicting fuel consumption and emissions for specific excavators and projects, which are clearly needed for confident evaluation and planning. This may at least partially explain why planners and contractors currently pay little attention to environmental factors when selecting equipment [17].

The aim of this study was to meet the need for a suitable technique for predicting the energy consumption and emissions of excavators during earth digging and haulage operations in infrastructure projects at planning stages, based on artificial neural networks (ANNs). The ANNs were developed using seven input parameters (excavator utilization rate, digging depth, cycle time, bucket payload, load factor, horsepower, and hauler capacity) and a target value for energy usage or CO_2 emissions per cubic meter earth handled. We also applied Discrete Event Simulations (DES) of earthmoving scenarios, in order to provide operational data for these activities that led to the generation of the final subsets of data required to build the proposed ANN models. This approach also allowed identification of effects of each of the input parameters on the target outputs, thereby providing further insights for planners when selecting machinery and planning earthmoving operations. In addition, a case study of a road construction project in Northern Sweden is used to demonstrate the applicability of the proposed ANN models.

1.1 Background for Using ANN in Construction Management

ANNs have a number of applications in construction management, including: predicting earthmoving operations [18]; computing an effectiveness ratio for designing earthworks [19]; computing optimum machinery settings [20]; determining site layout [21]; computing the duration of projects [22]; predicting the performance of contractors [23]; assessing the quality of projects [24]; improving pavement management systems [25]; identifying factors that affect labor productivity [26–28]; and selecting and managing the use of cranes [29]. ANN models for construction scheduling and cost optimization have been developed [30–34] by adapting the general neural dynamics model of Adeli and Park [35]. ANNs have also been used for cost estimation [36–47], predicting the productivity of construction equipment [47–50], and estimating overheads [51, 52]. However, ANNs have not been used, to our knowledge, to predict the environmental impact of construction projects.

2 Methodology

The objective of the study was to propose ANN models that can be used during the early planning stages of road and infrastructure projects to predict energy use and CO_2 emissions of excavators. The methodology is divided into three stages: the acquisition of information about excavators and haulers; the development of a model to simulate earthworks operations and thereby estimate the energy consumption and CO_2 emissions of different excavators under different operating conditions; and the training and testing of ANNs. The first two stages are used to generate the dataset with which we train and test the ANNs. Haulers are used to remove material from an excavation site; the number of haulers used and their respective characteristics can affect the utilization time of the excavators, thus having a bearing on energy consumption. Our overall methodology is summarized in the integrated definition for function modeling (IDEF0) framework presented in Fig. 1. In Table 1, we list the input and output parameters used in our framework.

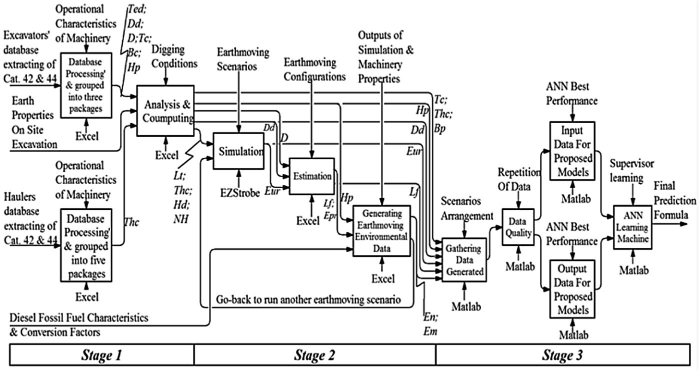

Excavators' database extracting of Cat. 42 & 44
Operational Characteristics of Machinery
Database Processing' & grouped into three packages
Excel
Earth Properties On Site Excavation
Ted; Dd; D;Tc; Bc; Hp
Digging Conditions
Analysis & Coumputing
Excel
Haulers database extracting of Cat. 42 & 44
Operational Characteristics of Machinery
Database Processing' & grouped into five packages
Excel
Thc
Lt; Thc; Hd; NH
Earthmoving Scenarios
Simulation
EZStrobe
Eur
Dd
D
Earthmoving Configurations
Estimation
Excel
Lf; Epr
Hp
Outputs of Simulation & Machinery Properties
Generating Earthmoving Environmental Data
Excel
Go-back to run another earthmoving scenario
Diesel Fossil Fuel Characteristics & Conversion Factors
Tc; Thc; Bp
Hp
Dd
Eur
Lf
En; Em
Scenarios Arrangement
Gathering Data Generated
Matlab
Repetition Of Data
Data Quality
Matlab
ANN Best Performance
Input Data For Proposed Models
Matlab
ANN Best Performance
Output Data For Proposed Models
Matlab
Supervisor learning
ANN Learning Machine
Matlab
Final Prediction Formula
Stage 1
Stage 2
Stage 3

Fig. 1. IDEF0 framework for excavators' energy consumption and CO_2 emission.

Table 1. Input parameters for the earthmoving scenarios.

Parameter	Notation	Units	Input to *Lf*?	Input to DES?	Input to *En* and *Em* estimation?	Input to ANN?
Type of earth digging	*Ted*	–	Yes			
Digging depth	*Dd*	M				Yes
Bank density	*D*	kg/m^3	Yes			
Excavator cycle time	*Tc*	min		Yes		Yes
Excavator bucket size	*Bc*	m^3				
Bucket payload	*Bp*	m^3				Yes
Horsepower	*Hp*	kW			Yes	Yes
Hauler heaped capacity	*Thc*	m^3		Yes		Yes
Loading time for each hauler capacity	*Lt*	min		Yes		
Hauling distance	*Hd*	km		Yes		
Number of haulers	*NH*	Integer		Yes		
Excavator utilization rate	*Eur*	decimal			Yes	Yes
Load factor	*Lf*	decimal			Yes	Yes
Excavator productivity	*Epr*	m^3/h			Yes	
Energy from each earthmoving scenario	*En*	MJ/m^3				Yes
Emission from each earthmoving scenario	*Em*	kg/m^3				Yes
Artificial neural networks	ANN	–				

2.1 Data Extraction

We obtained information about the operational characteristics of a number of excavators manufactured by Caterpillar from the relevant handbooks [53, 54]. These characteristics include: optimum cycle time for excavating different types of earth at maximum depth, bucket capacity, maximum engine horsepower, and haulage capacity. Cycle time refers to the time taken to excavate material, swing the bucket to a hauler, unload contents of the bucket onto the hauler and return to the excavation point. We also gathered information about the operational characteristics of a number of haulers manufactured by Caterpillar. This information is relevant to the simulation of earthwork operations, as haulers will be required to remove material that has been excavated. In particular, the capacity of a hauler and its speed are likely to have an effect on the utilization rate of an excavator. At the end of this stage, an equipment database was arranged in a separate excel sheet based on plans proposed to mimic all earthmoving scenarios that might happen in real earthmoving operations.

2.2 Simulation Model Through Discrete Event Simulation

In order to generate training and testing data for our ANNs, we developed a model to simulate earthwork operations using the Ezstrobe tool for discrete event simulation [55]. Our model takes into account the density of the material to be excavated, the digging depth, the excavator cycle time, the bucket size of the excavator, and the horsepower of the excavator. Relevant data about densities for each earth digging is an important element when specifying the load factor of the excavator. Therefore, this data was estimated within a range of values for each earth type based on a logarithm expression shown by Jassim et al. 2017 [56]. The utilization rate of an excavator depends on the time taken for a hauler to remove and dump material extracted by the excavator. Accordingly, our model also takes into account the capacity, speed, round-trip distance and dumping time of a hauler (see Fig. 2). We considered different earthmoving scenarios and multiple configurations for each scenario, each configuration having different values for the various input parameters described above.

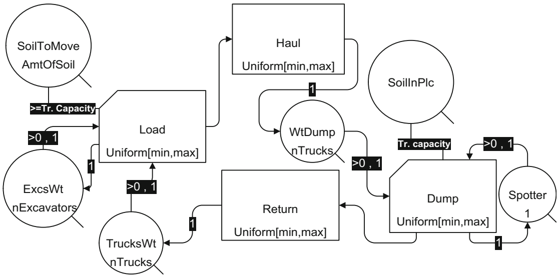


Fig. 2. Earthmoving template for operating excavators in different scenarios.

The excavators were divided into three groups of seven to nine models, based on their maximum digging or loading depth, and recommended type of earth excavation according to the Caterpillar handbooks. In each earthmoving scenario a single excavator was served with 2–9 haulers with optimal operating characteristics, i.e. highest compatibility with the excavator's capacity, and hauling distances.

The ultimate goal of our simulations is to compute an excavator's productivity rate at each level of utilization based on different earthmoving characteristics that use to estimate the energy consumption and CO_2 emissions per cubic meter of material excavated, which we call the normalized energy consumption and normalized CO_2 emissions, denoted *En* and *Em*, respectively. The most important output from the simulation, for our purposes, is the utilization rate of the excavator. The actual productivity of an excavator is based on its utilization rate, from which we compute an excavator's *En* and *Em* for a given job. In particular, we use formulas (1) – (5) below.

$$En = \left(\frac{SFC.Hp.Lf.E_{cf}}{\rho_{fuel}.Epr} \right); \tag{1}$$

$$Em = \left(\frac{SFC.Hp.Lf.E_{mcf}}{\rho_{fuel}.Epr}\right); \tag{2}$$

$$Lf = 0.0366e^{0.00136B_D}; \tag{3}$$

$$Epr = \left(\frac{Bp.Eur.60}{Tc}\right); \tag{4}$$

$$Eur = (1 - M_{wt}). \tag{5}$$

where *En* and *Em* are, respectively, the energy consumption (MJ/m^3) and CO$_2$ emissions (kg/m^3) of the excavator that was generated from each earthmoving scenario. *Lf* is the engine load factor (decimal), estimated using Eq. (3), based on prior work by Jassim et al. 2017 [56], where B_D represents the bank density of the material excavated (kg/m^3). *SFC* is the specific fuel consumption (0.22 kg/kW.h) to be set to a suitable value for engines with power in the range 28.8 to 370 kW [57]; H_p is the maximum design horsepower of the excavator engine (kW); E_{cf} is the conversion factor for the energy of each liter of diesel fuel (36 MJ/l) [58]; ρ_{fuel} is the specific gravity of the diesel fuel to be consumed (0.85 kg/l) [59], ranging between 0.83 and 0.87 kg/l; E_{mcf} is the conversion factor for the CO$_2$ of each liter of diesel fuel (2.6569 kg CO$_2$/l) [60]; *Epr* is a productivity rate (m^3/h) for each level of utilization to excavator; *Bp* is the bucket payload (m^3); and *Tc* is the cycle time (min). We compute *Eur* using Eq. (5), where M_{wt} is average content of queue of excavator (decimal).

2.3 Prediction Using ANN Models

We view *En* and *Em* as functions of excavator utilization rate, digging depth, cycle time, bucket payload, engine horsepower, load factor (derived from the density of the material being excavated) and hauler capacity. We use artificial neural nets (ANNs) to learn these functions, the results of our simulation providing the dataset with which we trained and tested our ANNs. Thus the inputs to the ANNs are excavator utilization rate, digging depth, cycle time, bucket payload, horsepower, load factor and hauler capacity; and the outputs are *En* and *Em*. Our earthmoving scenarios considered many different configurations, described in the previous section, each simulation characterized by a tuple of seven input variables and two output variables. We imported these tuples into Matlab and removed duplicates. The resulting dataset contained 19784 tuples with which we trained and tested our ANNs. The configurations considered in the simulations included all (15) permutations of three groups of excavators and five groups of 2-9 haulers with 24–60.2 m^3 capacity, used to move earth with 21 hauling distances (0.1 km and 0.5–10 km in 0.5 km increments). Each of these configurations was combined with excavation of earth with 15 densities and five digging depths.

The data processing inside an ANN model can be categorized into three phases. The first phase involves using training subset data to update the weight connections in the network layers using backward propagation at the training stage. The second phase, in parallel with the learning process, uses the testing subset to identify the responses of the designed neural network to the data that do not form part of the training data, but which are a part of the whole dataset and within its boundaries. In the third phase, the neural network utilizes data examples that do not belong to the other two subsets (i.e., training and testing) to produce a validation data subset that provides a final indication of model acceptability and validity. We conducted a number of trials in order to decide what proportion of our dataset to use for training our ANNs (the remainder to be used for testing). We considered values between 75% and 93%, leaving between 25% and 7% of the data, respectively, for testing. We found that using 80% of the data for training produced the best results. Perception Multilayer (PML) networks, a backward propagation learning method based on the Levenberg–Marquardt algorithm, were used for the training data in the neural network, using a sigmoid activation function [61].

The structure of an ANN, in terms of hidden nodes and layers, has a significant bearing on its ability to produce good results. Although a number of methods for determining the number of hidden nodes and layers have been suggested [62], none of them has gained widespread acceptance. Therefore, we used trial and error to determine the number of hidden nodes and layers. We found the ANN with the best performance, in terms of mean square error (MSE) and value correlation coefficient (R), had one hidden layer containing 15 nodes (see Fig. 3).

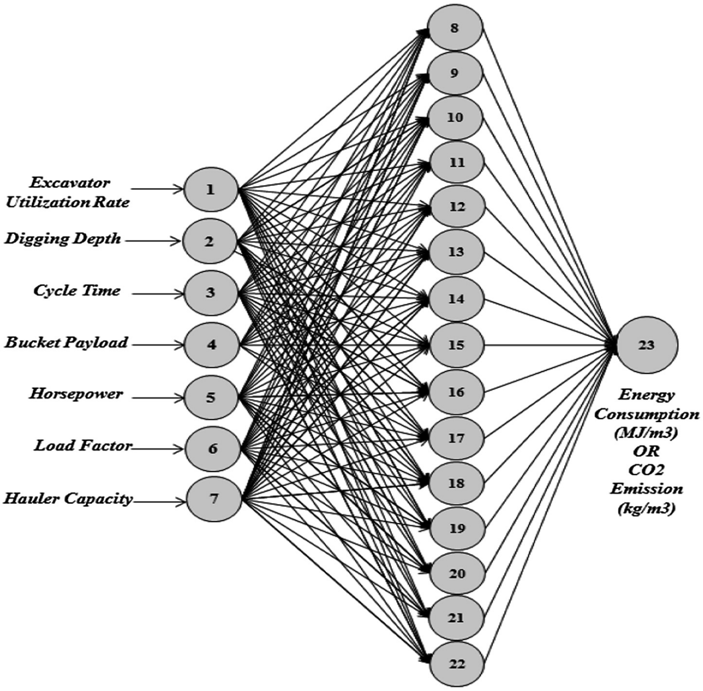


Fig. 3. An optimal architectural structure for the ANN proposed model.

All inputs and outputs were scaled to a value in the range [0,1], the default scaling of the sigmoid function used in the backward propagation learning algorithms [61]. In fact, the study used a scale for the data within the ranges [0.1, 0.9], in order to avoid slow learning rates for data points at the endpoints of this range [61], we scaled the inputs using the following formula:

$$X_s = \left(\frac{0.8}{Delta_i}\right)x_i + \left(0.9 - \frac{0.8x_{max}}{Delta_i}\right) \tag{6}$$

where X_s represent the scaled value of input parameter, x_i represents the normal value of input parameter, x_{max} represent the maximum value of input parameter (see Table 2), $Delta_i$ represent the amount of different between maximum and minimum value for each input parameter (see Table 2).

$$Y_r = \left(\frac{delta_o}{0.8}\right)y_s - 0.9\left(\frac{Delta_o}{0.8}\right) + y_{imax} \tag{7}$$

Table 2. Values used to scale input and output variables.

Delta value	***b1***	***b2***	***b3***	***b4***	***b5***	***b6***	***b7***	***yen***	***ym***	$\boldsymbol{\theta_y}$
	0.7882	4.1	0.18	3.2708	315	0.78	35.1	22.1338	1.6335	-0.17838
Maximum value	***b1***	***b2***	***b3***	***b4***	***b5***	***b6***	***b7***	***yenmax***	***ymmax***	$\boldsymbol{\theta_{yy}}$
	0.999	5.6	0.35	3.6316	382	0.9	60.2	22.5928	1.6674	–3.03634

where Y_r represent the final rescaled (i.e. normal) value for energy consumption or CO_2 emissions output by the ANN model, y_s represent the scaled value of output parameter, y_{imax} represent the maximum value of the output parameter for each ANN model, $Delta_o$ represent the difference between the maximum and minimum values for each output parameter. Table 2 shows the values $Delta$, y, and x_i used for the inputs to and outputs from our ANNs.

2.4 Sensitivity Analysis

In general, there are five reasons to implement sensitivity analysis in mathematical models [63]. We considered two of these reasons: the relative importance of each input parameter on the output [64]; and the effects of changing a particular input parameter on the output [63]. We used the partitioning weights method [65] to determine the relative importance to the various input parameters on the outputs, as used by Goh [66]. In addition, a study of the effects of changing input on output was achieved through the ANN models was done by Matlab.

3 Results of ANN Model Development

The results of our study can be divided into three parts: the ability of our ANNs to predict energy consumption and CO_2 emissions by excavators; the parameters of the ANNs; and the sensitivity analysis.

3.1 The Accuracy of Our ANNs

We used trial and error to determine (i) the relative sizes of the training and testing data subsets (ii) the number of hidden nodes, and (iii) the number of training cycles (epochs). We selected ANNs that minimized mean square error (MSE) and maximized the correlation coefficient value (R). Consequently, many trials with various sizes of data subsets and learning rates were carried out, and the best three trials for each ANN model are shown in Table 3.

Table 3. The values of different criterion to select optimum design for ANN predicting model, where N* = Design of ANN model; L* = learning rate; $S^{wholeData}$ = Size of whole data set; $S^{training}$ = Size of data subset training; $S^{testing}$ = Size of data subset testing; MSE = Mean square error for best training performance; Epochs = number of iteration to get on best output; $R^{training}$ = Correlation coefficient for output training data subsets (output vs. target); $R^{testing}$ = Correlation coefficient for output testing data subsets (output vs. target).

Item	N^*	L^*	$S^{wholeData}$	$S^{training}$	$S^{testing}$	MSE	Epochs	$R^{training}$	$R^{testing}$
Energy model	7-15-1	0.1	19874	15899	3975	$9.8731 * 10^{-6}$	45	0.99945	0.99930
	7-15-1	0.1	19874	15899	3975	$8.8629 * 10^{-6}$	40	0.99952	0.99945
	7-15-1	0.1	19874	15899	3975	$8.8121 * 10^{-6}$	57	0.99960	0.99930
Emission model	7-15-1	0.1	19874	15899	3975	$8.4165 * 10^{-6}$	75	0.99940	0.99931
	7-15-1	0.1	19874	15899	3975	$8.0091 * 10^{-6}$	103	0.99947	0.99930
	7-15-1	0.1	19874	15899	3975	$7.1063 * 10^{-6}$	31	0.99951	0.99940

We found the best ANN for predicting energy use had MSE of 8.8121×10^{-6} and R of 0.99960; and the best ANN for predicting CO_2 emissions had MSE of 7.1063×10^{-6} and R of 0.99951. Figures 4 and 5 illustrate how MSE varies with the number of epochs. The graphs in Figs. 6 and 7 compare actual results from our dataset with those predicted by our best ANNs. In the course of this study, we had to determine the number of hidden nodes in our ANNs. We found that ANNs having between $2n$ and $3n$ hidden nodes, where n is the number of input parameters, produced good results.

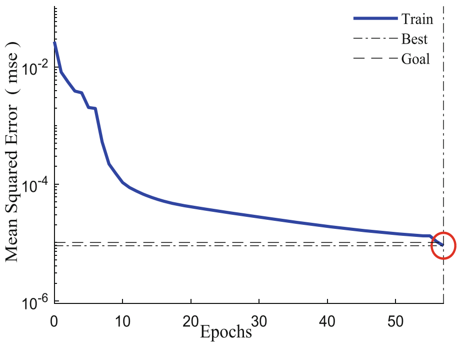


Fig. 4. Best training performance (8.8121e–06) of energy model.

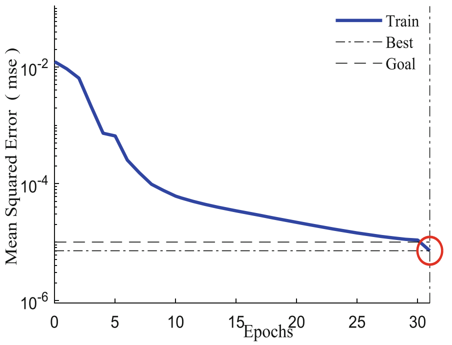


Fig. 5. Best training performance (7.1063e–06) of emission model.

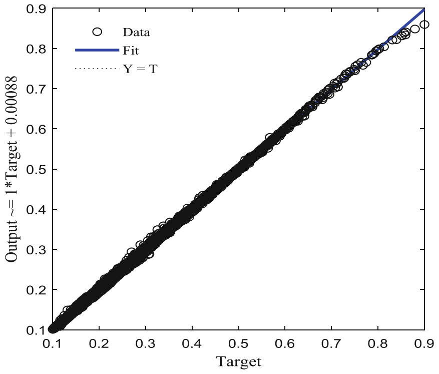


Fig. 6. Fitting for training data with R = 0.9996 of energy model.

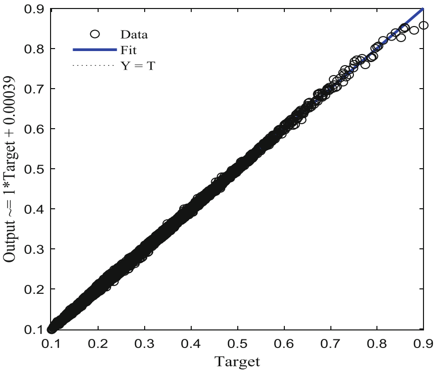


Fig. 7. Fitting for training data with R = 0.99951 of emission model.

3.2 Characteristics of the Best ANN Models

Our best ANN models for energy consumption and CO_2 emissions are characterized by equations and matrices (8–14, 22,24,26) and (15–21,23,25,27) respectively.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \tag{8}$$

$$D = A.B \tag{9}$$

$$E = D + C \tag{10}$$

$$F_i = [1./(1 + exp(-E))] \tag{11}$$

$$K = [f_1.h_1;f_2.h_2;f_3.h_3;f_4.h_4;f_5.h_5;\ldots\ldots\ldots\ldots;f_{13}.h_{13};f_{14}.h_{14};f_{15}.h_{15}] \tag{12}$$

$$S = \left[\sum\nolimits_{i=1}^{n} K_i\right] + \theta_y \tag{13}$$

$$E_{ncE} = [1./(1 + \exp(-S))] \tag{14}$$

$$\mathrm{AA} = \begin{bmatrix} \mathrm{aa}_{11} & \cdots & \mathrm{aa}_{1m} \\ \vdots & \ddots & \vdots \\ \mathrm{aa}_{n1} & \cdots & \mathrm{aa}_{nm} \end{bmatrix} \tag{15}$$

$$DD = AA.B \tag{16}$$

$$EE = DD + CC \tag{17}$$

$$FF_i = [1./(1 + exp(-EE))] \tag{18}$$

$$KK = [ff_1.hh_1;ff_2.hh_2;ff_3.hh_3;ff_4.hh_4;\ldots\ldots\ldots;ff_{14}.hh_{14};ff_{15}.hh_{15}] \tag{19}$$

$$SS = \left[\sum\nolimits_{i=1}^{n} KK_i\right] + \theta_{yy} \tag{20}$$

$$E_{mEco2} = [1./(1 + \exp(-SS))] \tag{21}$$

$$A = \begin{bmatrix}
1.932140848 & 0.803527591 & -1.782745417 & 1.924533029 & -2.660165535 & -0.738464227 & -0.011282961 \\
-2.450491914 & -0.584032721 & -8.517016308 & 1.203528269 & 0.316415599 & -2.214937557 & 0.049196998 \\
4.259578775 & -0.052615575 & -3.850580523 & 0.478384922 & -1.037926132 & 4.304888065 & -0.781231832 \\
-5.44121363 & -1.1952587 & -20.56220452 & 0.135356853 & -2.149573212 & -1.054825252 & 0.395045752 \\
-8.034040899 & -2.759345671 & 4.073301187 & -2.681670998 & 1.990615809 & 2.784141612 & -0.022095208 \\
5.223885048 & 1.535365025 & -24.07058253 & 0.880052266 & -0.466206059 & 0.565591519 & -0.105168183 \\
4.603686642 & 0.140184796 & 20.36752762 & 0.739017134 & -0.6282921 & 1.044233059 & -0.486854804 \\
-0.985925507 & 1.622920743 & -10.69330234 & 0.215578528 & 2.454459294 & 6.706365326 & -0.053253677 \\
-0.397422267 & 1.145133108 & -1.731305806 & -8.706794683 & 2.642496363 & 1.461554241 & -0.001479864 \\
-1.615045452 & -0.086527844 & -22.13453945 & -0.976616091 & -0.080359312 & 2.402065543 & 0.076298719 \\
3.520498674 & -0.655369432 & 11.49416631 & 0.459566248 & -0.055113145 & 5.123187205 & 0.216245321 \\
-0.172390839 & 0.057535095 & 4.914417475 & 1.054418545 & 0.094973667 & 7.878615009 & -0.345248181 \\
-3.463490871 & -0.46941335 & -22.2256822 & 1.41194256 & 1.14733147 & 0.199738963 & 0.194763093 \\
-0.448456756 & -0.394263557 & -27.44767221 & 0.527031406 & -1.052571231 & 0.495371301 & 0.140754062 \\
-0.894564559 & 0.404088378 & 0.554174236 & -1.397720452 & 1.094945691 & -4.419261116 & 0.00688263
\end{bmatrix} \tag{22}$$

$$AA = \begin{bmatrix}
4.867631174 & -0.397161498 & -8.786320838 & -0.498296452 & 0.022204655 & -4.124610753 & 0.178794524 \\
-5.091798071 & -1.21227655 & 13.61759692 & -1.737449188 & -2.730571483 & 0.660435471 & 0.138957494 \\
-0.333484018 & 0.219109321 & 2.586514755 & 0.608066658 & 1.211694026 & -1.077101818 & 0.068507745 \\
1.00982082 & -0.464720205 & -19.89394813 & 4.038128855 & -0.480997885 & -2.419712642 & -0.154497882 \\
-4.91889251 & -1.649302591 & 5.318060813 & -0.855815909 & 1.324306895 & -0.554374294 & 0.088987207 \\
1.499784125 & 0.491192863 & 23.90518949 & -1.407546998 & -0.133229215 & 2.43397077 & 0.333262754 \\
-1.302177291 & 0.978176977 & 22.64128341 & -0.720880286 & 0.300224905 & 0.007366332 & 0.701677941 \\
-1.389804943 & -0.067394971 & -31.09635797 & -0.622052714 & -0.143524164 & 4.524135673 & 0.186124297 \\
0.806921254 & -0.200229005 & -17.15515889 & -1.905544775 & 1.008250698 & -0.51900249 & -0.26826444 \\
-0.488178784 & 0.053462458 & -0.303813641 & -0.161877062 & 1.22382961 & -4.031223092 & 0.015201459 \\
-3.659105414 & 0.556444719 & -15.35996648 & 0.753689546 & -1.191746907 & 1.107354136 & -0.134685362 \\
-1.841427901 & -0.913453924 & 13.80018321 & 4.211316942 & -1.171086508 & 0.024109108 & 0.172702683 \\
-9.375571555 & 2.428295924 & -0.773126996 & -1.980528342 & 1.228766135 & 2.615901476 & -0.005179659 \\
-3.744876462 & 0.210610221 & -11.85723402 & 0.44169883 & 0.707917123 & -4.290128221 & -0.013577638 \\
1.10959389 & 0.034570516 & -0.007568218 & 5.043498902 & -0.832038908 & -2.381852199 & 0.050634168
\end{bmatrix} \tag{23}$$

$$C = \begin{bmatrix} 1.041766514 \\ 17.29667353 \\ -11.5507509 \\ 11.85241343 \\ -3.07060579 \\ 3.754147876 \\ -9.69767671 \\ 1.333712663 \\ -1.00290924 \\ -0.58613044 \\ -8.03728687 \\ 0.676223085 \\ 3.52829281 \\ 4.046076791 \\ -0.69521567 \end{bmatrix} \tag{24}$$

$$CC = \begin{bmatrix} -8.09626125 \\ -2.97492249 \\ 0.103705872 \\ 15.02836671 \\ -0.418224983 \\ -10.85503434 \\ 3.203194773 \\ 8.094233417 \\ 1.843729458 \\ -0.57303871 \\ 3.403818357 \\ -3.559810548 \\ -2.60872849 \\ 8.40409903 \\ 0.424143033 \end{bmatrix} \tag{25}$$

$$H = \begin{bmatrix} -2.569258585 \\ 2.113810825 \\ -1.300843046 \\ -0.400743852 \\ 8.242200556 \\ -0.213570597 \\ -0.290736535 \\ 0.148933449 \\ 3.617688423 \\ -1.323174558 \\ -0.097331803 \\ -1.051554462 \\ 0.439472123 \\ -0.553806803 \\ -3.26319513 \end{bmatrix} \tag{26}$$

$$HH = \begin{bmatrix} 2.197083407 \\ -0.784958615 \\ 5.54022482 \\ -0.114712438 \\ 1.5029456 \\ 0.043303659 \\ 1.903136058 \\ -0.142252555 \\ 0.93058256 \\ -3.588235108 \\ -0.444273514 \\ -1.010659608 \\ 5.864278474 \\ -0.077769708 \\ -4.222412425 \end{bmatrix} \tag{27}$$

The matrices *A* and *AA* represent the weight connection matrix between the input and hidden layers for energy and emission models, respectively. Matrix *B* represents the scaled values for the input parameters. Matrices *C* and *CC* represent the bias values (i.e., threshold) of nodes in the hidden layer for the energy and emission models, respectively. Matrices *D* and *DD* represent the matrices resulting from the multiplication of the weight connections and scaled input parameters matrices. Matrices *H* and *HH* represent the weights connection vector matrix between the hidden and output layers for the energy and emission models, respectively. *F* and *FF* represent the matrices resulting from applying a sigmoid function to each weight connection between the input and hidden layers. *K* and *KK* are vector matrices for elements facing each other in *F* and *H* and *FF* and *HH,* respectively. (Note that this step is not typical for matrix multiplication, but it is regarded as multiplication only for parallel elements in both of them). θ_y and θ_{yy} represent the bias values (i.e., threshold) of nodes in the output layer (see Table 2). *S* and *SS* represent the sum of the bias value of the node output layer and the sum of the elements in the *K* and *KK* matrices respectively. E_{ncE} and E_{mEco2} represent the predicted values of an excavator's energy consumption and CO_2 emissions (per cubic meter of material excavated), respectively.

3.3 Results of the Sensitivity Analysis

Figures 8 and 9 show the relative importance of each input parameter for energy consumption and CO_2 emissions, respectively. It is clear from these figures that excavator cycle time (*Tc*) is by far the most significant factor in determining energy consumption and CO_2 emissions. Load factor (*Lf*) and utilization rate (*Eur*) are also important, with bucket payload (*Bp*), horsepower (*Hp*), digging depth (*Dp*), and hauler capacity (*Thc*) having far less influence.

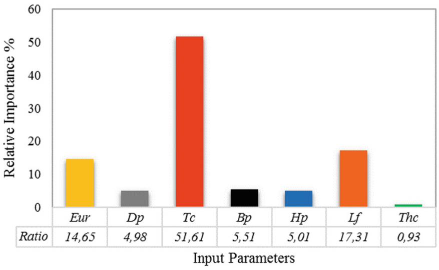


Fig. 8. Relative importance for ANN energy model.

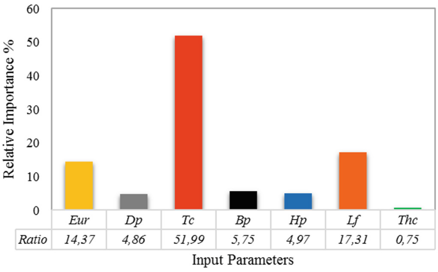


Fig. 9. Relative importance for ANN emission model.

In addition, Figs. 10, 11, 12, 13, 14 and 15 show how energy consumption varies with each input parameter. As one might expect, increasing cycle time, engine load factor, or horsepower increases the amount of energy consumed, whereas increasing (i.e. improving) utilization rate of an excavator can decrease the amount of energy consumed, per cubic meter excavated. Similar changes to these parameters cause similar effects on CO_2 emission because CO_2 emissions are directly proportional to fuel and energy consumed.

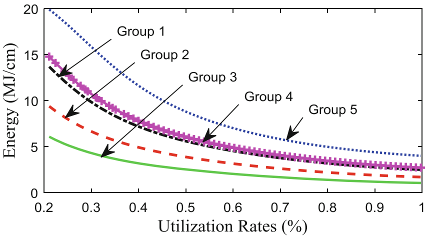


Fig. 10. Relationships between energy and utilization rate of excavator.

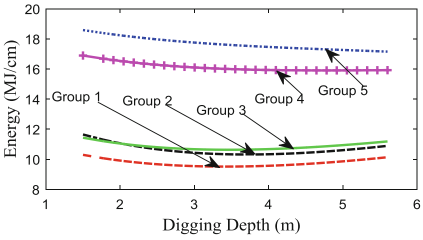


Fig. 11. Relationships between energy and digging depth of excavator.

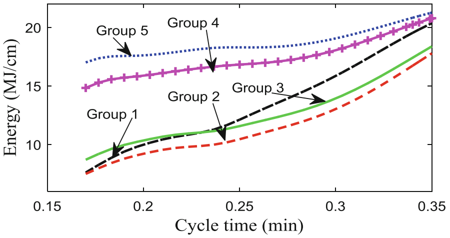


Fig. 12. Relationships between energy and cycle time of excavator.

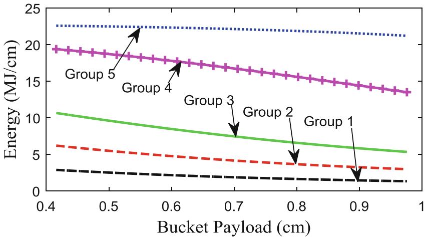


Fig. 13. Relationships between energy and bucket payload of excavator.

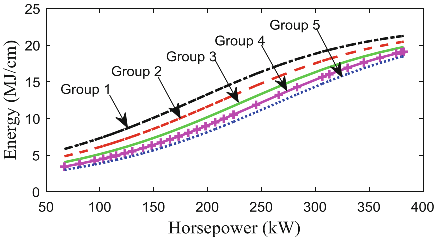


Fig. 14. Relationships between energy and horsepower of excavator.

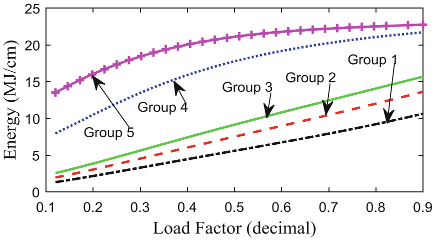


Fig. 15. Relationships between energy and engine load factor of excavator.

4 Application of the ANN Predicting Model in a Case Study

The environmental impact of a construction project is an important consideration, with incentives and rewards available for environmentally friendly projects [9]. This case study demonstrates how our ANNs can help planners predict the energy consumption and CO_2 emissions used by different types of excavators for a given project, and thus make an informed decision about which excavator(s) to use.

The case study concerns a new road construction project (väg 870) in Kiruna municipality, northern Sweden. We used the documentation supplied with the project tender to obtain information about the earthwork activities and operations involved, such as depth of excavations, volume of material to be excavated, and size and location of borrow pits and disposal areas.

We considered three sections of the road and three different depths of excavation within each section. Each layer was assumed to have a different density (increasing with depth), with densities varying between sections. We computed the load factor for each layer at each section, as shown in Table 4. We considered three different excavator models (312D, M315D and 324D); the horsepower and bucket payload of each excavator are given in Table 5. We assume that a particular model of hauler (773G) will be used and that the utilization rate of each excavator will be 60% throughout.

Table 4. Characteristics of sections and layers in the case study.

Section	Layer	Bank Density ((kg/m^3)	Volume (m^3)	Load factor (decimal)
A	I	1900	900	0.49
	II	1950	1100	0.52
	III	2020	550	0.57
B	I	1550	900	0.30
	II	1750	1100	0.39
	III	1890	550	0.48
C	I	1150	900	0.18
	II	1425	1100	0.25
	III	1660	550	0.35

Table 5. Characteristics of excavators used in the case study.

Excavator	Bucket payload (m^3)	Horsepower (kW)	Cycle time (min)
312D	0.68	67	0.17–0.32
M315D	0.98	101	0.17–0.21
324D	1.31	140	0.17–0.26

However, energy use and emissions for each layer were computed using Eq. 28, with energy use and emissions from each section computed using Eq. 29, based on the quantity of materials required to be excavated from each layer.

$$E_L = E_{ANN} * Q_L \tag{28}$$

$$E_S = \sum\nolimits_{i=1}^{n} E_{L_i} \tag{29}$$

where E_L is the energy or emission value from hauling materials of a specific depth layer, E_{ANN} is the normalized (rescaled) value for energy consumed or CO_2 emitted per cubic meter of earth excavated, according to the proposed ANN models, Q_L is the quantity of materials hauled from a specific depth layer. E_S is the total amount of energy consumed, or CO_2 emitted, when excavating all the materials in each section, $i = 1,2,\ 3,\ \ldots,\ n$; where n = number of layers in each section.

We computed predicted values for E_{ncE} and E_{mEco2} (as described in Subsect. 3.2) and multiplied these values by the volume of material to be excavated. Table 6 shows the resulting estimates for total energy consumption and CO_2 emissions.

Our results suggest that the M315D excavator would be the most suitable for this project (see Fig. 16). We note that the 312D model was predicted to consume more energy than the M315D, despite having a smaller engine. This can be attributed to the fact that some of the properties of the earthworks required by the project lie outside the ranges recommended for this model, and confirms the results of the sensitivity analysis in Sect. 3.

Table 6. Predicted energy consumption and CO_2 emissions.

Excavator	Section	Layer	Energy (MJ/m^3)	Emission (kg/m^3)	Total energy per layer (MJ)	Total CO_2 emissions per layer (kg)	Total energy per section (MJ)	Total CO_2 emissions per section (kg)
312D	A	I	2.21794	0.167811	1996.2	151.0	9875	886.6
		II	4.00552	0.27303	4406.1	300.3		
		III	6.31398	0.791338	3472.7	435.2		
	B	I	0.80747	0.046753	726.7	42.08	6892	646.9
		II	2.9713	0.202615	3268.4	222.9		
		III	5.26639	0.694473	2896.5	381.9		
	C	I	0.75137	0.042094	676.2	37.9	4804	477.1
		II	1.84918	0.124997	2034.1	137.5		
		III	3.80615	0.548592	2093.4	301.7		
M315D	A	I	0.93157	0.097195	838.4	87.48	5865	413.1
		II	2.81918	0.176848	3101.1	194.5		
		III	3.50078	0.238257	1925.4	131.0		
	B	I	0.60389	0.035852	543.5	32.27	4453	276.5
		II	2.14771	0.128505	2362.5	141.4		
		III	2.8136	0.18715	1547.5	102.9		
	C	I	0.34109	0.003797	306.9	3.42	2984	164.2
		II	1.34237	0.073924	1476. 7	81.32		
		III	2.18319	0.144438	1200.8	79.44		
324D	A	I	0.80747	0.046753	726.7	42.08	7965	499.8
		II	3.39596	0.196488	3735.6	216.1		
		III	4.69996	0.314958	2584.9	173.2		
	B	I	1.14695	0.068495	1032.3	61.65	5977	362.2
		II	2.63379	0.147526	2897.2	162.3		
		III	3.72263	0.251548	2047.4	138.4		
	C	I	0.6474	0.027947	582.7	25.15	4037	230.4
		II	1.69024	0.088234	1859.3	97.06		
		III	2.90041	0.196777	1595.2	108.2		

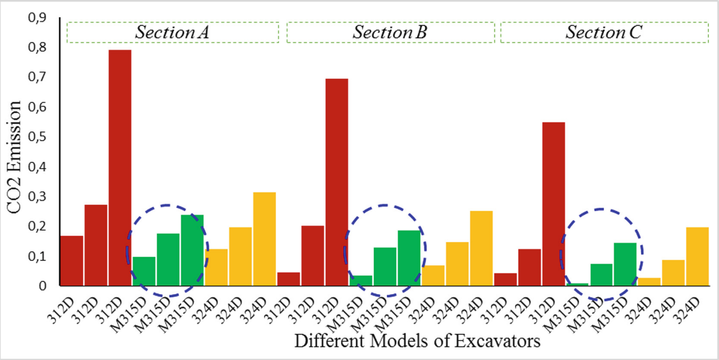


Fig. 16. CO_2 emission values for three models of excavators.

5 Discussion

We used discrete event simulation to generate synthetic data associated with earthworks operations. Little actual data has been gathered from such operations because of the complexity and cost of acquiring measurements in the field. Therefore, we believe the dataset generated by our simulation provides a useful resource for future research that applies machine learning techniques, such as ANNs, to problems associated with earthworks operations.

In our simulation we followed prior work in this area [55, 67] and used a uniform distribution to model the variation in the loading time for a given combination of excavator and hauler. We did try using other distribution functions, such as a triangle distribution, but this had little effect on the outputs of the simulation, such as utilization rates.

ANNs are widely used to estimate the output of a complex, non-linear function whose precise structure is unknown. We previously showed how an ANN can be used to predict energy consumption and CO_2 emissions [56]. In that earlier work and the work reported in this paper, we determined the best ANN by computing relevant statistics (MSE and R) from the outputs produced by the ANN and the corresponding outputs in the training and testing datasets.

Of all the factors affecting the energy consumption (and thus CO_2 emissions) of an excavator, cycle time is the most important [16, 68, 69], especially when the excavator is used outside the ranges recommended by the manufacturer. For example, Fig. 12 shows that the energy consumed by excavators in group 1 (comprising the smallest excavators expected to operate with short cycle times) is predicted to increase disproportionately compared with that consumed by other excavators as cycle time increases. This result suggests the proposed model is able to identify unsuitable machinery selection for each area of work. In Fig. 15, energy use for different engine sizes (horsepower) showed a positive relationship with load factor where all other factors in the ANN model remained constant. However, the excavators with large engines (i.e. those in groups 4 and 5) showed a rapid increase in energy consumption against load factor when the load factor was less than 40%, indicating that such excavators should not be used to haul materials with a bank density of less than (1750 kg/m^3) or to excavate at less than half the depth recommended by the manufacturer. The first point is reinforced by the recommendation to use machinery in groups 4 and 5 to excavate hard clay materials [16]. The utilization rate of an excavator can be enhanced by reducing total idle time (e.g. waiting time) of the machinery, which can be achieved by selecting the best earthmoving fleet to work with the excavator. Our results in Fig. 10 show that a reduction in the amount of energy used (per cubic meter excavated) of between 30% to 40% can be achieved by improving excavator utilization rate from 40% to 60%.

Figure 11 shows that digging depth has little effect on energy consumption and CO_2 emissions. The figure also confirms that energy consumption is minimized when an excavator operates at the digging depth recommended in the manufacturer's handbook (as seen in the curves for groups 1, 2 and 3) [54]. We would expect that an increase in the engine size of an excavator, in terms of horsepower, would increase the amount of energy consumed. Figure 14 confirms this expected behavior.

The effect of bucket payload can be addressed in terms of fill factors for each range of materials density per each hauling cycle, which is also influenced by the operators' experiences. For example, in Fig. 13, this parameter was tested for five models of excavator under similar conditions such as utilization rate, digging depth, cycle time, and load factor, but across different horsepower figures. Unsurprisingly, increasing the bucket payload - the amount of material excavated by the excavator in a single cycle - decreases the amount of energy consumed (per cubic meter excavated).

In order to gain more insight into the results of relative importance for the input parameters, we compared our results with the results from a previous study [56], noting that bucket payload was the fifth most important factor in the original study and the fourth most important in this study. This result may be due to the influence of additional parameters used in the current study and the different situations for the data examined in the two studies. Where the previous study assumed the same utilization rate for all earthmoving scenarios for forecasting energy consumption and CO_2 emissions, the current study considered a number of different utilization rates. Moreover, the current study considered energy consumption per cubic meter of material excavated, whereas the earlier study measured energy consumption per hour. Finally, the difference between the sizes of the datasets used in the two studies may have resulted in the different outcomes. In fact, the size of the dataset in this study is approximately four times greater than the original study thereby for the larger dataset in this study is likely to lead to more robust results. Thus. this is another demonstration of the efficiency and sensitivity of the ANN models to tackle complex and non-linear relationships for input parameters.

The results of our case study suggest the M315D excavator would be the most suitable one for the proposed project. We note that the 312D model, despite having a smaller engine size, was predicted to consume more energy than the M315D model. This can be attributed to the fact that some of the properties of the earthworks required by the project, such as digging depth, lie outside the ranges recommended for the 312D model in the manufacturer's handbook.

Therefore, the selection of suitable machinery to earthworks operations can improve earthmoving performance of both sides where can help planners and construction managers to reduce emissions and save money that is mainly coming of fuel cost.

6 Conclusion

We have shown that DES provides a good way to generate synthetic data with which to model the energy consumption and CO_2 emissions of excavators in a variety of earthworks scenarios. The resulting dataset may prove to be a valuable resource for subsequent research that uses machine learning to solve problems associated with earthworks operations. We have also shown that ANNs can be used by the construction industry to predict, at the planning stage of a project, the energy use and CO_2 emissions of excavators. Thus, ANN models can guide the selection of excavators for a given project, based on environmental factors. This study suggests the two most important factors affecting energy consumption are cycle time and engine load factor, confirming the results of a previous study [56], even though the two studies used different units to measure energy consumption. Our sensitivity analysis demonstrates the effect of individual parameters on energy consumption and CO_2 emissions, thereby enhancing our understanding of the potential environmental impacts of using different excavators under different conditions. Moreover, the analysis provides important insights into how best to make use of excavators while minimizing their environmental impact. In particular, increasing the utilization rate of excavators leads to a significant reduction in energy consumption (per cubic meter of material excavated), which suggests that appropriate choices for haulers and locations of dumping sites can play an important part in reducing the environmental impact of earthworks. Our analysis also helps us to identify excavators that would be unsuitable, in terms of energy consumption, for particular kinds of earthworks.

References

1. Anair, D.: Digging up trouble: the health risks of construction pollution in California. Union of Concerned Scientists (2006)
2. NSW EPA: Reducing emissions from non-road diesel engines: An information report prepared for the NSW EPA. ISBN:9781743597316, EPA 2014/0 586 (2014)
3. Des Participants, Liste.: Good practice guidance and uncertainty management in national greenhouse gas inventories. Order (2001)
4. Hoel, M., Kverndokk, S.: Depletion of fossil fuels and the impacts of global warming. Resour. Energy Econ. **18**, 115–136 (1996)
5. Avetisyan, H.G., Miller-Hooks, E., Melanta, S.: Decision models to support greenhouse gas emissions reduction from transportation construction projects. J. Constr. Eng. Manage. **138**, 631–641 (2011)
6. Dallmann T., Menon A.: Technology pathways for diesel engines used in non-road vehicles and equipment. Int. Counc. Clean Transp. (ICCT) (2016)
7. Sandanayake, M., Zhang, G., Setunge, S., Thomas, C.M.: Environmental emissions of construction equipment usage in pile foundation construction process—a case study. In: Shen, L., Ye, K., Mao, C. (eds.) Proceedings of the 19th International Symposium on Advancement of Construction Management and Real Estate, pp. 327–339. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46994-1_28
8. Keijzer, E., Leegwater, G., de Vos-Effting, S., de Wit, M.: Carbon footprint comparison of innovative techniques in the construction and maintenance of road infrastructure in The netherlands. Environ. Sci. Policy **54**, 218–225 (2015)
9. Trafikverket: The swedish transport administration's efforts for improving energy efficiency and for climate mitigation (2012)
10. EPA: MOVES2014a: Latest version of MOtor vehicle emission simulator (MOVES), united states environmental protection agency. EPA EPA-420-F-15-046 (2014)
11. CARB: California air resources board; In-use off-road equipment - 2011 inventory model. California Air Resources Board (2011). http://www.arb.ca.gov/msei/categories.htm. Accessed 05 Mar 2016
12. Hajji, A.M., Lewis, P.: Development of productivity-based estimating tool for energy and air emissions from earthwork construction activities. Smart Sustain. Built Environ. **2**, 84–100 (2013)
13. Hajji, A.M., Lewis, M.P.: Development of productivity-based estimating tool for fuel use and emissions from earthwork construction activities (2013)

14. Melanta, S., Miller-Hooks, E., Avetisyan, H.G.: Carbon footprint estimation tool for transportation construction projects. J. Constr. Eng. Manage. **139**, 547–555 (2012)
15. Clark, N.N., Kern, J.M., Atkinson, C.M., Nine, R.D.: Factors affecting heavy-duty diesel vehicle emissions. J. Air Waste Manage. Assoc. **52**, 84–94 (2002)
16. Barati, K., Shen, X.: Emissions modelling of earthmoving equipment. vol. 33, no. 1 (2016)
17. Hajji A.M., Muladi, Larasati A.: 'ENPROD' MODEL–estimating the energy impact of the use of heavy duty construction equipment by using productivity rate. vol. 1778, p. 030008 (2016)
18. Shi, J.J.: A neural network based system for predicting earthmoving production. Constr. Manag. Econ. **17**, 463–471 (1999)
19. Schabowicz, K., Hoła, B.: Application of artificial neural networks in predicting earthmoving machinery effectiveness ratios. Arch. Civil Mech. Eng. **8**, 73–84 (2008)
20. Hola, B., Schabowicz, K.: Estimation of earthworks execution time cost by means of artificial neural networks. Autom. Constr. **19**, 570–579 (2010)
21. Yeh, I.: Construction-site layout using annealed neural network. J. Comput. Civ. Eng. **9**, 201–208 (1995)
22. Bhokha, S., Ogunlana, S.O.: Application of artificial neural network to forecast construction duration of buildings at the predesign stage. Eng. Constr. Archit. Manag. **6**, 133–144 (1999)
23. Cheung, S.O., Wong, P.S.P., Fung, A.S., Coffey, W.: Predicting project performance through neural networks. Int. J. Proj. Manage. **24**, 207–215 (2006)
24. Shi, H.: Application of unascertained method and neural networks to quality assessment of construction project. vol. 1, pp. 52–55 (2009)
25. Amin, M.S.R., Amador-Jiménez, L.E.: Pavement management with dynamic traffic and artificial neural network: A case study of montreal. Can. J. Civil Eng. **43**, 241–251 (2015)
26. Portas, J., AbouRizk, S.: Neural network model for estimating construction productivity. J. Constr. Eng. Manage. **123**, 399–410 (1997)
27. Heravi, G., Eslamdoost, E.: Applying artificial neural networks for measuring and predicting construction-labor productivity. J. Constr. Eng. Manage. **141**, 04015032 (2015)
28. Muqeem S., Idrus A.B., Khamidi, M.F., Zakaria, M.S.: Prediction modelling of construction labor production rates using artificial neural network, pp. 32–36 (2011)
29. Sawhney, A., Mund, A.: IntelliCranes: an integrated crane type and model selection system. Constr. Manag. Econ. **19**, 227–237 (2001)
30. Luu, V.T., Kim, S.: Neural network model for construction cost prediction of apartment projects in vietnam. Korean J. Constr. Eng. Manag. **10**, 139–147 (2009)
31. Hua, G.B.: Residential construction demand forecasting using economic indicators: a comparative study of artificial neural networks and multiple regression. Constr. Manage. Econ. **14**, 25–34 (1996)
32. Roxas C.L.C., Ongpeng J.M.C.: An artificial neural network approach to structural cost estimation of building projects in the Philippines (2014)
33. Gulcicek, U., Ozkan, O., Gunduz, M., Demir, I.H.: Cost assessment of construction projects through neural networks. Canadian J. Civ. Eng. **40**, 574–579 (2013)
34. Arafa, M., Alqedra, M.: Early stage cost estimation of buildings construction projects using artificial neural networks. J. Artif. Intell. **4**, 63–75 (2011)
35. Adeli, H., Karim, A.: Scheduling/cost optimization and neural dynamics model for construction. J. Constr. Eng. Manag. **123**, 450–458 (1997)
36. Günaydın, H.M., Doğan, S.Z.: A neural network approach for early cost estimation of structural systems of buildings. Int. J. Project Manag. **22**, 595–602 (2004)
37. Hegazy, T., Ayed, A.: Neural network model for parametric cost estimation of highway projects. J. Constr. Eng. Manag. **124**, 210–218 (1998)

38. Kim, G., Yoon, J., An, S., Cho, H., Kang, K.: Neural network model incorporating a genetic algorithm in estimating construction costs. Build. Environ. **39**, 1333–1340 (2004)
39. Xiao-chen D.: Application of neural network in the cost estimation of highway engineering. J. Comput. **5**(11), 1763 (2010)
40. Elhag, T., Boussabaine, A.: An artificial neural system for cost estimation of construction projects (1998)
41. Emsley, M.W., Lowe, D.J., Duff, A.R., Harding, A., Hickson, A.: Data modelling and the application of a neural network approach to the prediction of total construction costs. Constr. Manag. Econ. **20**, 465–472 (2002)
42. Iranmanesh, S.H., Zarezadeh, M.: Application of artificial neural network to forecast actual cost of a project to improve earned value management system. pp. 240–243 (2008)
43. Sodikov, J.: Cost estimation of highway projects in developing countries: Artificial neural network approach. J. East. Asia Soc. Transp. Stud. **6**, 1036–1047 (2005)
44. Williams, T.P.: Predicting changes in construction cost indexes using neural networks. J. Constr. Eng. Manage. **120**, 306–320 (1994)
45. Wilmot, C.G., Mei, B.: Neural network modeling of highway construction costs. J. Constr. Eng. Manage. **131**, 765–771 (2005)
46. Zhao Y., Chua D.K.: Relationship between productivity and non value-adding activities (2003)
47. Boussabaine, A.H.: A neural network system for productivity forecasting. pp. 375–381 (1995)
48. Moselhi, O., Hegazy, T., Fazio, P.: Potential applications of neural networks in construction. Can. J. Civ. Eng. **19**, 521–529 (1992)
49. Tam, C., Tong, T.K., Tse, S.L.: Artificial neural networks model for predicting excavator productivity. Eng. Constr. Archit. Manag. **9**, 446–452 (2002)
50. Ok, S.C., Sinha, S.K.: Construction equipment productivity estimation using artificial neural network model. Constr. Manage. Econ. **24**, 1029–1044 (2006)
51. Assaf, S.A., Bubshait, A.A., Atiyah, S., Al-Shahri, M.: The management of construction company overhead costs. Int. J. Proj. Manag. **19**, 295–303 (2001)
52. ElSawy I., Hosny H., Razek M.A.: A neural network model for construction projects site overhead cost estimating in Egypt (2011)
53. Cat. 44: Caterpillar performance handbook (2015)
54. Cat. 42: Caterpillar performance handbook (2012)
55. Martinez, J.C.: EZStrobe: general-purpose simulation system based on activity cycle diagrams. pp. 1556–1564 (2001)
56. Jassim, H.S., Lu, W., Olofsson, T.: Predicting energy consumption and CO2 emissions of excavators in earthwork operations: an artificial neural network model. Sustainability **9**, 1257 (2017)
57. Klanfar, M., Korman, T., Kujundžić, T.: Fuel consumption and engine load factors of equipment in quarrying of crushed stone. Teh. vjesn. **23**, 163–169 (2016)
58. ICLEI: Clean fleets guide. procuring clean and efficient road vehicles; international council for local environmental initiatives (ICLEI-europe) local governments for sustainability: Freiburg, Germany, ICLEI-Europe (2014)
59. Zhang, S., Wu, Y., Liu, H., Huang, R., Yang, L., Li, Z., Fu, L., Hao, J.: Real-world fuel consumption and CO2 emissions of urban public buses in beijing. Appl. Energy **113**, 1645–1655 (2014)
60. DECC Defra: Guidelines to defra/DECC's GHG conversion factors for company reporting (2011)
61. Oreta, A.W.C.: Simulating size effect on shear strength of RC beams without stirrups using neural networks. Eng. Struct. **26**, 681–691 (2004)

62. Cross, S.S., Harrison, R.F., Kennedy, R.L.: Introduction to neural networks. Lancet **346**, 1075–1079 (1995)
63. Hamby, D.: A review of techniques for parameter sensitivity analysis of environmental models. Environ. Monit. Assess. **32**, 135–154 (1994)
64. Helton, J., Iman, R., Johnson, J., Leigh, C.: Uncertainty and sensitivity analysis of a model for multicomponent aerosol dynamics. Nucl. Technol. **73**, 320–342 (1986)
65. Garson, D.G.: Interpreting neural network connection weights. Artif. Intell. Expert **6**, 47–51 (1991)
66. Goh, A.: Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. **9**, 143–151 (1995)
67. Martínez, J.C.: EZStrobe—General-purpose simulation system based on activity cycle diagrams. pp. 341–348 (1998)
68. FUEL ECONOMY: Fuelling savings in tough times - blutip power technologies. pp. 38–46 (2015)
69. Komatsu: Komatsu specification and application handbook (2009)

Life-Cycle Design Support

Intelligent Computing for Building Performance Analysis

Pieter de Wilde(✉)

Plymouth University, Plymouth, PL4 8AA, UK
pieter.dewilde@plymouth.ac.uk

Abstract. A challenge towards the intelligent use of computing in civil and architectural engineering is the definition of the questions that the ICT technology has to address. To some extent this is implicitly covered by activities such as the definition of search and option spaces, development of model views, the specification of objective functions, definitions of ontologies, or the development of multi-criterion decision methods. However, the underlying needs and drivers of design, construction and facility management processes of buildings are hard to capture, while they are essential to effective use of computing techniques. This paper reviews the starting point for intelligent computing within the domain of building performance analysis. It explores how approaches from the field of requirement engineering may help to support proper definition of computational needs, while embedding computational analysis efforts within the wider context of assessment approaches that are available in the building domain.

Keywords: Building performance · Requirement engineering
Intelligent computing

1 Introduction

The field of engineering informatics covers, amongst others, the use of computing technology in the building and construction discipline. A solid body of knowledge about engineering informatics has been developed, as represented by articles in peer-reviewed academic journals such as *Advanced Engineering Informatics*, books like the *BIM Handbook* [1] or *Fundamentals of Computer-Aided Engineering* [2] and, indeed, the proceedings of the annual workshop of the European Group for Intelligent Computing in Engineering (EG-ICE). Topics covered within engineering informatics cover issues such as data management, optimization and search, visualization, machine learning, (webbased) collaboration, building information modelling, and many others.

A challenge to intelligent use of computing in civil and architectural engineering is the definition of the questions that the ICT technology has to address. These questions are the true drivers of the computing effort and are highly important in discerning between intelligent and not-so-intelligent use of ICT. However, their definition is often left implicit. To some extent they are covered by activities such as the definition of search and option spaces, development of model views, the specification of objective functions, definitions of ontologies, or the development of multi-criterion decision methods. But

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 457–471, 2018.
https://doi.org/10.1007/978-3-319-91635-4_23

the underlying needs and drivers of design, construction and facility management processes of buildings are hard to capture, while they are essential to effective use of computing approaches.

Building Performance Analysis is a wide field that deals with three constituent parts: an engineering view that explorers how well buildings meet functional requirements, a process view that studies the performance of the building construction process, and an aesthetic view that covers the architectural performance of buildings. In further detail, the engineering view deals with building quality, workload capacity, resource saving, timeliness and responsiveness [3]. Typical performance aspects of interest include structural stability, processing capacity, fire safety, energy efficiency, thermal comfort, lighting levels, acoustical comfort, and indoor environmental quality. The activity of building performance analysis takes place along the full building design life cycle, from initial definition of need, via design, actual construction and commissioning, management and operation, refurbishment and retrofit, to end-of-life disposal of buildings and their constituent parts. It brings together a wide array of stakeholders, including client, architect, civil engineers with a range of specialisms, contractors, facility managers and others. Combined with the fact that most buildings are complex, bespoke products that represent a system of systems, and that many performance aspects interact, this creates a challenging situation where the starting point for computational analysis effort is highly unique.

The aim of this paper is to explore the definition of underlying questions and drivers that from the starting point for intelligent computing within the domain of building performance analysis. The following objectives have been identified:

Objective 1: review the background of current building performance analysis software and systems;
Objective 2: investigate structured software analysis approaches from other domains;
Objective 3: develop suggestions on how drivers for building performance analysis may be captured, and explore how this may support the evolution of new approaches in the computational assessment of building performance;

The paper is builds on an extensive review of literature on building performance analysis that has been conducted in the context of a forthcoming book on the subject of Building Performance Analysis [4], but now taking the subject specifically into the computing domain and this extending the scope of the discussion. It positions intelligent computing of building performance in a wider context that compares and contrast computational analysis – mainly in the guise of building performance simulation – with other assessment approaches such as physical measurement, stakeholder assessment, and expert judgment.

2 Background on Building Performance Analysis Software Systems Development

Building performance analysis has a long history. Even in primitive shelters as constructed by early humans performance aspects like protecting occupants from the elements and providing safety from wild animals plays a key role. Initially performance of buildings will have been explored through trial and error. Later it became the domain of ‘master builders’ and architects. A seminal contribution to the field are the books on architecture by the Roman architect Vitruvius – also a civil engineer – who considered that buildings must possess three key qualities: *firmitas, utilitas* and *venustas*, roughly translated as strength, utility and beauty [5]. Further disciplines within the building domain emerged during the industrial revolution, which saw the development of specialisms like structural engineering and building services engineering. Interestingly, this still seems to impact the situation today. While different aspects all fit the notion of building performance, there still is a clear split between the domain of structural engineering and a different ‘blood group’ of aspects that are clustered as ‘building science’; building science typically covers heat and mass transfer, lighting, acoustics, and indoor air quality. A further body of knowledge on building performance was established in the late 1960s to early 1980s, spearheaded by the work of the Building Performance Research Unit (BPRU) at the University of Strathclyde and the CIB Working Commission W60, with the latter focussing on ‘working with the performance approach to building’. Pressures from issues like sustainability, climate change, limited resources, health and safety and occupant wellbeing keep moving the field forward. However, it must be noted that thermal aspects (energy use, thermal comfort) seem to be rather dominant in the computing side of the building science domain, followed by some interest in lighting and acoustics. Other aspects such as building evacuation modeling are niche areas, while some areas like burglary resistance lack meaningful simulation approaches.

Computing has always played an important role in quantifying the performance of a building for the key performance aspects. The introduction of desktop computing in the mid 1970s saw a step change in possibilities. This led to the emergence of a new domain named building performance simulation. For the history of this field see for instance the descriptions by Augenbroe [6] or Clarke [7]; for the detailed history of a selection of whole-building thermal simulation tools see for instance the work by Oh and Haberl [8]. It is interesting to note that the building science area has tended to develop ‘closed tool boxes’ in the form of programs like DOE-2, ESP-r, TRNSYS and EnergyPlus, while the structural engineering area has shown a tendency to stay closer to general engineering approaches and underlying mathematical equations as available in programs like Matlab, ANSYS and Mathematica.

Evolution of building performance simulation tools is a slow process. Many tools used today have been in existence for years: TRNSYS dates back to 1973, DOE-2 to 1975, and ESP-r to 1974. EnergyPlus was launched in its first version as recent as 1997 but includes a legacy of DOE-2 and BLAST. Even with regular updates, as are provided for EnergyPlus, this means that many of the underlying assumptions and computational routines have been around for a long time. For comparison, the more general tools have similar histories: ANSYS was first released in 1971, Matlab in 1984, and Mathematica in 1988.

A lot of recent efforts in development of tools for building performance analysis seems to be invested in building shells around existent building performance analysis engines, such as DesignBuilder, Safaira and OpenStudio environments around EnergyPlus, or IES around Apache.

The evolution of building performance analysis software cannot be seen isolated from the development of Building Information Models or BIM. Seminal work in the area as described by Eastman [9] shows how BIM systems emerged partly as a result of the desire to share data amongst various computer applications, in what was then known as product models. Work like the EU Combine Project [10] shows that these efforts specifically aimed for performance areas such as building energy efficiency, costs and aesthetics. Over the years, two types of developments can be observed: the development of tool-independent BIM infrastructure, such as the BuildingSmart International Foundation Classes (IFC), and the development of relatively 'closed' suites of interoperable tools such as those connected within the IES Software system. Obviously, the tool-independent infrastructure allows for flexibility and thus sees wider application, whereas the closed interoperable systems have a more constrained use domain. However, the tool-independent approach tends to suffer from information overload. This has led to the development of domain-specific filters, named Model View Definitions (MVDs). While MVDs are a step towards better management, they are not yet perfect; Lee et al. discuss the challenges in defining MVDs and how these issues may lead to inconsistencies in exchange specifications [11]. An overview of some of the practical issues when using BIM in a multi-disciplinary collaboration is provided by Singh et al. [12]. The 'closed' suites like DesignBuilder or IES limit the number of interactions and hence are less prone to data exchange problems, but this comes at the price of reduced flexibility. However, IES at the moment seems to be a well-accepted industry standard in the building services engineering sector; both IES and DesignBuilder enjoy a wide uptake in the education sector.

While the literature offers a wide description of technical details of building performance analysis tools – for instance the EnergyPlus Engineering Reference document alone is 847 pages long – there is a sparsity of information on the user needs that these tools aim to address. There are several academic papers, such as Petersen and Svendsen [13] or Negendahl [14] that describe the needs that tools need to respond to, but these do not provide insights into the requirements that drive actual tool development by the major software houses and academic communities. Papers from tool developers tend to focus on the features that their tools provide, rather than on the underlying requirements.

A general understanding of what existing tools respond to can be achieved by looking at the various tool capabilities that are used to define tool categories in the Building Energy Software Tool Directory (maintained by IBPSA-USA), available from www.buildingenergysoftwaretools.com, as presented in Table 1.

The following observations can be made with regards to these categories. First of all, while the name of the directory singles out the single performance aspect of energy, the list includes tools that deal with other aspects such as lighting and water. Secondly, while most categories relate to professional building design and facility management, training is recognized as a separate need. Thirdly, the categories of calibration and weather data analysis in fact deal with modeling efforts, rather than straight building performance analysis tasks. Fourth these categories mix analysis activities, such as load prediction and auditing, with a building system typology, such as the split between envelope and HVAC systems.

Another interesting issue is that of the intended users of building performance analysis software. The Building Energy Software Tool Directory lists the target "audience" for each individual tool, naming for instance: architects, architectural designers, architectural engineers, builders, building energy modelers, consultants, contractors, daylighting designers, design evaluators, educators, energy code writers, energy managers, engineers, homeowners, HVAC designers, lighting designers, managers, manufacturers, mechanical engineers, policy makers, professionals, researchers, simulation experts, students, sustainable design engineers, tenants and urban designers. Obviously this is a very wide range of tool users, with a strong variation in background in terms of proficiency in using computational tools as well as training in the principles that underlie building performance analysis.

While the academic literature only offers a limited insight in the user needs that the existent tools try to meet, a generally accepted approach appears to be to use BIM systems to define building properties, pushing data from the BIM to a range of analysis tools, and then proceeding with specific evaluations. One of the many examples that describes this approach is the work by Oduyemi and Okoroh [15], which lists the following steps being required: (i) description of the site (ii) description of the building (iii) selection of relevant 'design indicators' (iv) development of baseline performance levels (v) exploration of '*what-if*' scenarios which focus on specific interest, such as system and operational parameters and (vi) uncertainty and sensitivity analysis. The notion of '*what-if analysis*' seems to be a wider trend in the industry, and rests on the premise that the best way to support building performance analysis is to build a model of a building, and then explore the impact of the variation of properties and parameters; see for instance Hopfe and Hensen [16]. More advanced approaches are available from the domain of statistics, where the theory of 'design of experiments' (DOE) employs the principles of randomization, replication and blocking in order to allow for factorial experimentation with efficient evaluation of the variation across a set of different parameters [17]. This approach can become more demanding in situations where the analyst may want to explore different system configurations, such as in the case of selection of HVAC system components, or building retrofit: in some cases there may be pre-configured system models that can be switched on and off, but in cases where a system needs to be introduced from scratch it may require a significant modelling effort.

Within the building performance analysis field, and especially those sub-areas that are concerned with environmental and sustainability issues, there is special attention for the use of computational tools during design. The use of tools to support design decisions has important advantages: since the building is not yet in existence, this is the only way to predict the performance of what during design are merely plans. It allows to subject design proposals to exactly identical testing conditions, something that is often very hard to do in real life buildings, where the best one may achieve is a semi-controlled experiment, since one has only limited control over things like occupant behaviour and climate conditions.

There is deep debate how to best support building design, with some authors emphasizing the need to equip designers with tacit knowledge [18] while others suggest further

Table 1. Tool capability categories used by the building energy software tools directory.

Whole-building energy simulation	Building energy benchmarking
Load calculations	Lighting simulation
HVAC system selection and sizing	Indoor air quality simulation
Parametrics and optimization	Life-cycle analysis
Model input calibration	Detailed envelope simulation
Energy conservation measures	Detailed component simulation
Code compliance	Solar and photovoltaic analysis
Ratings and certificates	Electrical system simulation
Utility bill and meter data analysis	Water use analysis
Weather data and climate analysis	Training services
Building energy auditing	Other

rationalization of the design process and stress the importance of design decisions that are underpinned with evidence [19]. Further attention is directed to attempts to match computational efforts to characteristics of the design process, leading to a longstanding and persistent claim that tools need to support fast design evolution and permutation [20]. Similarly, there is significant discussion about the need to support early design decisions; it is generally believed that early design decisions may have more leverage on achievement of building performance, whereas later decisions concern a building design that is already 'locked in place' and thus have less impact. This conflicts with the amount of detail that may be used in the evaluation of performance, leading to what is commonly known as the 'design paradox' [21, 22]. Further work promotes the combination of building performance analysis tools with optimization algorithms as a way forward to arrive at optimal design solutions [23, 24].

While these are all serious issues, it leads to a situation where the criteria for performance analysis remain very generic. For instance, Attia et al. list the following issues as important in getting simulation tools integrated into the design process: (i) quick analysis that supports decision making (ii) incorporation of uncertainty and sensitivity analysis of key parameters (iii) capability for weather analysis and suggestion of appropriate solutions (iv) ability to be used across various design stages [25]. Such recommendations are useful at a holistic level, but are hardly appropriate to guide the creation of software in a way that can be validated and verified.

There are two efforts that strictly speaking are not drivers for software development, but which aim to provide a better fit between design activities and computational analysis efforts. However, these provide some interesting insights in the deeper requirements. The first is the notion of 'performance assessment methods' or PAMs, developed in the context of International Energy Agency Annex 21 which ran from 1988 to 1993. A PAM sets out the information needs for a particular building performance analysis effort, such as the analysis of overheating risk. The idea behind PAMs was to provide a tool-independent definition of analysis needs in order to enable comparison of calculation methods. In theory, this might also be used as a specification of requirements for tool development [26]. The second is the development of Analysis Functions (AFs) in the context of the Design Analysis Interface (DAI) Initiative. AFs are defined to enable an efficient data parsing from a central BIM model to dedicated performance analysis tools, providing a template of the information that is required to undertake a specific assessment. AFs are linked to tasks in a process that is modelled and enacted in a workflow management system; however the stakeholders and their activities used are rather traditional and based on academic insights rather than operational studies of actual needs [27].

This brief review of the background of building performance analysis software systems gives an overview of main developments and trends. Obviously there is more detail in various papers and handbooks that introduce the tools currently on the market. However, it seems safe to state that most building performance analysis tools have been developed iteratively, using a trail and error process, and that natural selection over a period of around 50 years has resulted in the emergence and retention of the present toolset. Current tools are still heavily reliant on legacy 'calculation engines', with the models embedded in many tools dating back to the 1970s; recent efforts seem to focus more on building shells around tools, with those shells offering advanced interfaces for handling building geometry, quick access to default systems and settings, and reducing modeling requirements – see for instance the efforts on DesignBuilder, Sefaira and OpenStudio around the EnergyPlus simulation engine, or IES around the legacy Apache engine. A special branch of software development is also emerging around the Rhinoceros 3D design application, where a set of add-on applications such as Grasshopper, Ladybug and Honeybee helps to interface with simulation engines like EnergyPlus and Radiance, as well as use generative functions.

3 Structured Approaches from Other Domains

While the development of building performance analysis software requires deep subject knowledge, efforts in this area sometimes become rather inward-looking, ignoring developments in the wider context. For instance there is a significant body of knowledge on Software Engineering, as exemplified by the seminal textbook by Pressman [28]. Typically, this stresses the fact that all software development takes place in response to some kind of business demand. Any programming efforts are preceded by identification of the stakeholders and their needs, followed by planning of the process, design of the system architecture, software construction/coding, testing and ultimately deployment. There are different processes, which can be highly linear and prescriptive, incremental and iterative, or evolutionary. Yet, as stated by Pressman: '*Understanding the requirements of a problem is among the most difficult tasks that face a software engineer..... even if customers and end-users are explicit in their needs, those needs will change throughout the project. Requirements engineering is hard.*' Yet this is the essential work that defines how software fits in a business process, meets the need of the client, and how end-users will interact with the product.

Software development can be seen as part of the wider realm of Systems Engineering [29]. Systems engineering is an interdisciplinary field of science that deals with the design and management of systems, where systems may be physical systems ('hardware'), IT systems ('software'), business and services. Like Pressman on software, the International Council on Systems Engineering (INCOSE) emphasizes the need to start with the customer, saying that systems engineering '*focusses on defining customer needs and required functionality early in the development cycle, documenting requirements, and then proceeding with design synthesis and system validation while considering the complete problem*' [ibid]. This requires the analysis of the business or mission, definition of stakeholder needs and requirements, and identification of system requirements. Gilb points out that a key challenge in project management and engineering efforts is that it is difficult to articulate requirements and to cope with changes in requirements that have been set [30]. He goes on to list the following list of key issues for proper definition of requirements:

- identification of critical stakeholders;
- separation of requirements from design ideas (keeping requirements and solutions apart);
- prioritization of key requirements that are critical for system success;
- definition of what will be considered system success, or system failure;
- comparison of requirements to benchmarks;
- development of timescales for delivery.

While these are relevant issues to keep in mind, it still leaves open how one actually does identify the stakeholder needs. This can be done through requirement engineering, the branch of systems engineering that aims to capture and describe the client needs and expectations for a new product. Requirement Engineering consist of the following fundamental activities [29]:

(1) requirement elicitation, the process of identifying stakeholders in a new product and their needs and desires;
(2) requirement documentation, which involves the description of the requirements stemming from the elicitation process in words and models;
(3) validation and negotiation, the process of checking that the requirements are complete, reflect true stakeholder needs, and solving any conflicts;
(4) requirement management, maintaining track of changes and ensuring consistency.

Further detail on requirements engineering can be found in the publications by Pohl and Rupp [31] or Robertson and Robertson [32]. Amongst others, these works suggest the use of various techniques like brainstorming, analysis of existing systems to identify stakeholders, workshops and interviews to elicit requirements from stakeholders, and expression of requirements by means of language template ('boilerplates') and formal modeling techniques. A special technique is the definition of 'use cases', which describe the detailed interaction between a user and a system. Use cases help to reduce the complexity of large systems, dividing the overall system functionality into smaller views that correspond with the activities that system users will undertake when employing the system. Modern visualization languages such as UML (Unified Modeling Language) and IDEF (Integration DEFintion) have dedicated diagrams that help to depict use cases.

Some further interesting observations can be made from within the construction/ building engineering domain. For instance, Lucas et al. demonstrate the use of UML Use Cases to explore the information needs of a healthcare facility, and how this may be used to design an IT system for facility management of a hospital [33]. Wang et al.

explore the anticipated activities of future building occupants and how this relates to the building lay-out using a BIM-based system; interestingly this is almost the development of use cases but for the building itself, not for the software used to support the design process [34]. While these efforts focus on facility design and management, it shows the applicability of UML Use Cases in the construction sector. Girodon et al. point out that software may help to automate repetitive design tasks, but that efficient systems should relate to expertise and knowledge of their users; they suggest an agent-based approach to tailor systems to specific users [35]. Their approach not only discerns different actors and activities but also different roles and missions. Chong and Chen discuss that stakeholder (customer) needs are not static, but may evolve and change; their customer requirement analysis and forecast (CRAF) system attempts to address this challenge [36]. Dynamic requirement development is definitely something one would expect for software engineering, where regular updates are now common place across most platforms. Wang et al. go on to explore how user requirements may change due to interactions between the product, user behaviour, motivation, and perceived value [34].

Golzarpoor et al. note that modern information systems combine data management with the application of efficient and effective processes. However, in building and construction the emphasis in IT systems is on product information; process control and workflow management receive only very limited attention [37]. By way of example Luo et al. present a system for developing and managing the functional performance specifications for a building, focussing on support of the briefing process rather than on requirements posed for the building performance analysis software [38]. A general structure for supporting IT-based collaboration between project partners is presented by Ren et al. However, the focus in this work is on supporting the planning process; it shows the complexity of the many interactions between various actors, roles and processes but does not delve into the specific tasks that are to be supported by building performance analysis software [39].

Most interest in building performance analysis computations goes towards quantitative assessment. Providing qualitative design support to architects and engineers however is not straightforward either due to the complexity of buildings, uncertainties, and information often being vague and incomplete at design stage [40]. However, the starting point for any intelligent computational effort should be a clear requirement definition. For instance, in mechanical engineering, design space exploration is explicitly linked to the definition of a system architecture. Gadeyne et al. [41] provide an overview of the description of the design space for gearbox design using the Object Constraint Language (OCL) and Systems Modeling Language (SysML). Even with a such a well-defined system, with limited degrees of freedom, capturing the design space is clearly non-trivial.

4 Definition and Modeling of Building Performance Analysis Drivers

Taking a steer from requirements engineering, this section provides an initial inroad into the definition and modeling of building performance analysis efforts. A first step is the exploration of stakeholders in computational analysis, and definition of use cases.

The range of stakeholders that have an interest in building performance is long. Many authors have provided overviews, listing clients, developers, building occupants or users, government at local and national level, society at large, architects, engineers, specialist consultants, contractors, product manufacturers, facility managers, financial institutes, insurers, and others. Amongst these, two groups get the most interest where it comes to definition of building performance analysis tools: architects and engineers; see for instance the papers by Bleil de Souza [18] and Attia [25]. However, it must be born in mind that these stakeholders are in fact categories and risk stereotyping. In practice buildings are mostly designed and engineered by many people with a wide range of personal traits, expertise, training and qualifications. In terms of qualifications one may discern between architects, architectural engineers, architectural technologists, mechanical engineers, building services engineers, construction managers, building science consultants, energy specialists, all with their own professional bodies (for instance in the UK these would be the likes of the Royal Institute of British Architects RIBA, Chartered Institute of Architectural Technologists CIAT, the Institution of Mechanical Engineers ImechE, Chartered Institution of Building Services Engineers CIBSE, or the Energy Institute IE). Even when using these formal qualifiers, real life will depend on personal interactions and dynamics; for instance Negendahl points out three main cases: (i) an architect working with an engineer as assistant (ii) a hybrid professional that combines both roles in one person and (iii) an architect and engineer working as equal partners [14]. Obviously this may be expanded with a case that fits situations where technology is dominant, such as chemical plants, where the case would be (iv) an engineering taking the main lead, with an architect as assistant. These same models can be applied to all other permutations of professionals, and expanded to teams that involve more than just two actors. Further complexity is added by the fact that these professional qualifications can come with different levels of training; for instance one may distinguish between novices, intermediate-level and experts in each category.

A proper identification of the many stakeholders in Building Performance Analysis however is just the first step in proper defining what is required of computational efforts. Further work is needed to identify which of these stakeholders, or what group of stakeholders, may be the software system user. This then can be followed up by an attempt to identify use cases. A first step in this direction that can be found in the literature are the building simulation use patterns as described by Tucker and Bleil de Souza [42]. This exploration of patterns could be expanded, empirically exploring the current activities of a range of AEC professional and how studying why and how they use existent systems. This would lead to a range of UML Use Case diagrams, as illustrated by Fig. 1, which is based on a theoretical range of software uses by an Architectural Engineer.

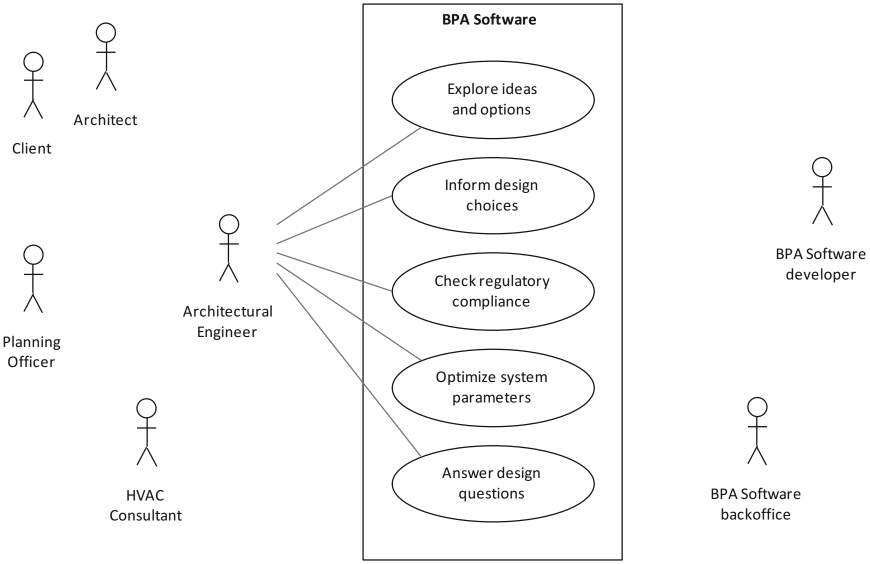


Fig. 1. Simple UML use case diagram depicting the use of building performance analysis software by an architectural engineer.

However, the situations depicted in Fig. 1 and in typical Use Case diagrams are only a first starting point for the interaction between any stakeholder and typical building performance analysis software. Further analysis will reveal a range of tasks and activities, any underlying process logic, and the interdependency of various data streams. Again as an illustration, Fig. 2 depicts a range of typical activities encountered when

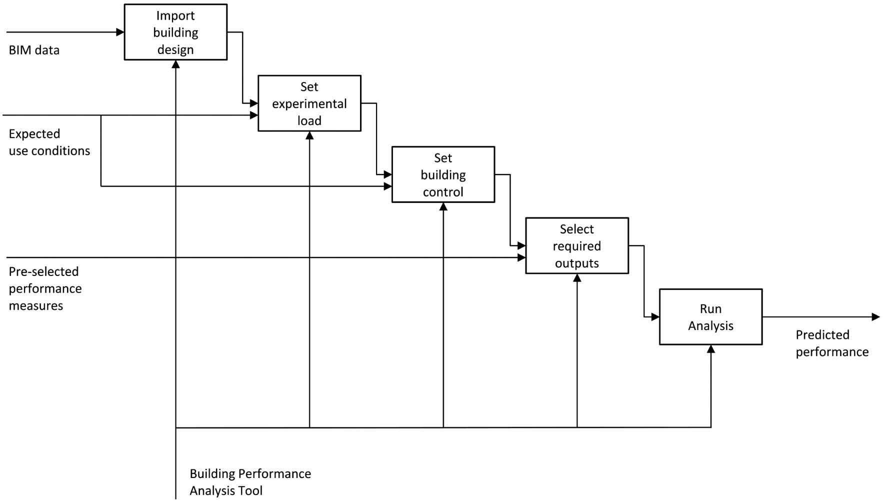


Fig. 2. Building performance analysis process logic (starting-point/use case dependent)

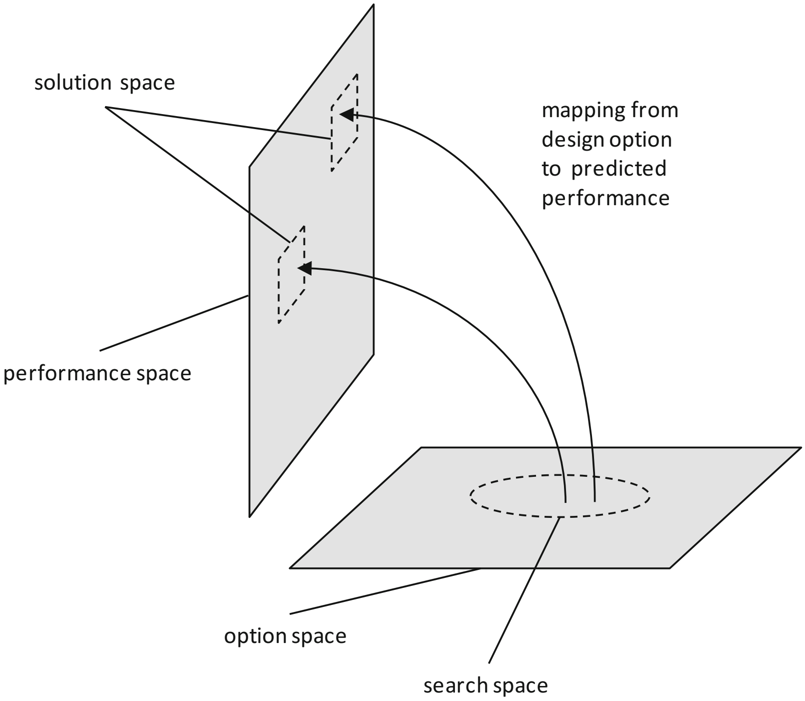


Fig. 3. Contextualization of mapping from search space to solution space.

using the current generation of software, such as EnergyPlus or DesignBuilder software. Note that this is only a crude attempt at high-level identification of tasks. Extensive empirical research would be needed to review actual workflows as occurring in the daily practice of various software users, after which categorization and standardization would provide realistic views in more detail. However, the depicted workflow already is helpful in defining various tasks needed to capitalize on the information that can be gained from a typical BIM system, as well as additional data needed to set up more specific analysis requests. At actual task level the workflow will be highly specific, providing a further focus for approaches that go further than the still quite generic Model View Definitions.

Beyond the workflow logic as captured in Fig. 2, further exploration is needed with respect to two terms that are used rather generally in literature on building performance analysis efforts: the ‘search space’ and ‘solution space’. Typically these terms are encountered in the context of optimization. Search space is used to describe all system variants that are considered, whereas solution space is used to describe the corresponding predictions of system performance. However, further detail and discussion can be added. On the system side, one could make a difference between the ‘option space’ and the ‘search space’. Here the option space would contain all the systems variants that are theoretically possible; the search space would be a subset of the option space, only containing those elements that are of interest to the stakeholder. Note that option spaces and search spaces may be continuous or discrete. For buildings, they are typically huge, with a large number of design parameters, system options, and attributes for each system. For each option, there then is a another space that contains the performance of that system option. Again, this is a multidimensional space, containing performance for aspects such as thermal comfort, energy efficiency, structural stability, visual comfort, and many others. As the list of performance aspects is long, it is unlikely that the whole performance space will be studied; instead, actual analysis efforts will focus on subsets which are limited by prediction capabilities and efforts, which can be designated as the solution space. The task of building performance analysis software then is to provide a mapping from search space(s) to solution space(s). See Fig. 3. Note that further research on option, search, performance and solution spaces is also required. On the system side, there are deep constraints in terms of system dimensional coordination, system compatibility, and standardization that render option and search spaces far from simple and continuous. On the performance and solution space side, single deterministic performance values are a typical oversimplification, and work is needed to incorporate various uncertainties and sensitivities in the mapping.

5 Discussion and Conclusion

This paper has provided an overview of the background of the software currently available for building performance analysis. It has also explored some of the approaches available in other domains, notable systems engineering and requirements engineering, which may help set more specific software development aims. This is followed by an exploration of how some concepts may be applied in the building performance analysis domain, and how these may lead to novel insights.

The paper concludes that intelligent computing of building performance has much to gain from investing time in proper development of analysis needs, in order to ensure that efforts are directed towards the things that really matter to the stakeholders. Rather than continue along a path of slow evolution, where the needs that drive the use of computational tools are left implicit, this approach would drive more specific and intelligent deployment of software. This may lead to a step change in the development of building performance analysis software that is more responsive to analysis needs, and is less dependent on tacit knowledge of software uses about the potential of their analysis tools. This paper present a first exploration towards further efforts in this direction.

References

1. Eastman, C., Teicholz, P., Sacks, R., Liston, K.: BIM Handbook – A Guide to Building Information Modelling, 2nd edn. Wiley, Hoboken (2011)
2. Raphael, B., Smith, I.: Fundamentals of Computer-Aided Engineering. Wiley, Chichester (2003)
3. de Wilde, P.: The concept of building performance in building performance simulation – a critical review. In: Barnaby, C., Wetter, M. (eds.) 15th International IBPSA Conference on Building Simulation 2017, San Francisco (2017)
4. de Wilde, P.: Building Performance Analysis. Wiley-Blackwell, Hoboken (2018)

5. Morgan, M.: Vitruvius: The Ten Books on Architecture. Dover Publications, New York (1960)
6. Augenbroe, G.: Trends in building simulation. In: Malkawi, A., Augenbroe, G. (eds.) Advanced Building Simulation. Spon Press, New York (2003)
7. Clarke, J.: Energy Simulation in Building Design, 2nd edn. Butterworth-Heinemann, Oxford (2001)
8. Oh, S., Haberl, J.: Origins of analysis methods used to design high-performance commercial buildings: whole-building energy simulation. Sci. Technol. Built Environ. **22**(1), 118–137 (2016)
9. Eastman, C.: Building Product Models: Computer Environments Supporting Design and Construction. CRC Press, Boca Raton (1999)
10. Augenbroe, G.: COMBINE 2 Final Report. Commission of the European Communities, Brussels (1995)
11. Lee, Y., Eastman, C., Solihin, W.: An ontology-based approach for developing data exchange requirements and model views of building information modeling. Adv. Eng. Inform. **30**, 354–367 (2016)
12. Singh, V., Gu, N., Wang, X.: A theoretical framework of a BIM-based multi-disciplinary collaboration platform. Autom. Constr. **20**, 134–144 (2011)
13. Petersen, S., Svendsen, S.: Method and simulation programs informed decisions in the early stages of building design. Energy Build. **42**, 1113–1119 (2010)
14. Negendahl, K.: Building performance simulation in the early design stage: an introduction to integrated dynamic models. Autom. Constr. **54**, 39–53 (2015)
15. Oduyemi, O., Okoroh, M.: Building performance modelling for sustainable building design. Int. J. Sustain. Built Environ. **5**, 461–469 (2016)
16. Hopfe, C., Hensen, J.: Uncertainty analysis in building performance simulation for design support. Energy Build. **43**, 2798–2805 (2011)
17. Montgomery, D.: Design and Analysis of Experiments, 8th edn. Wiley, Hoboken (2013)
18. de Souza, C.B.: Contrasting paradigms of design thinking: the building thermal simulation tool user vs the building designer. Autom. Constr. **22**, 112–122 (2012)
19. Becker, R.: Fundamentals of performance-based building design. Build. Simul. **1**(4), 356–371 (2008)
20. Marsh, A.: Peformance Analysis and Conceptual Design. Ph.D. thesis. University of Western Australia, Perth (1997)
21. Macdonald, I: Quantifying the Effects of Uncertainty in Building Simulation. Ph.D. thesis. University of Strathclyde, Glasgow (2002)
22. Marsh, R.: LCA profiles for building components: strategies for the early design process. Build. Res. Inf. **44**(4), 358–375 (2016)
23. Machairas, V., Tsangrassoulis, A., Axarli, K.: Algorithms for optimization of building design: a review. Renew. Sustain. Energy Rev. **31**, 101–112 (2014)
24. Nguyen, A., Reiter, S., Rigo, P.: A review on simulation-based optimization methods applied to building performance analysis. Appl. Energy **113**, 1043–1058 (2014)
25. Attia, S., Hensen, J., Beltrán, L., De Herde, A.: Selection criteria for building performance simulation tools: contrasting architects' and engineers' needs. J. Build. Perform. Simul. **5**(3), 155–169 (2012)
26. International Energy Agency: Annex 21 – Calculation of Energy and Environmental Performance of Buildings. Subtask B: Appropriate Use of Programs. Volume 1: Executive Summary. Building Research Establishment, Watford (1994)

27. Augenbroe, G., de Wilde, P., Moon, Y., Malkawi, A., Choudhary, R., Mahdavi, A., Brame, R.: Design Analysis Interface (DAI) Final Report. Georgia Institute of Technology, Atlanta (2003)
28. Pressman, R.: Software Engineering: A Practitioner's Approach, 6th edn. McGraw-Hill, New York (2005)
29. INCOSE: Systems Engineering Handbook: a Guide for System Life Cycle Processes and Activities. 3th edn. Wiley, Hoboken (2015)
30. Gilb, T.: Competitive Engineering: a Handbook for Systems Engineering, Requirements Engineering, and Software Engineering using Planguage. Butterworth-Heinemann, Oxford (2005)
31. Pohl, K., Rupp, C.: Requirements Engineering Fundamentals, 2nd edn. RockyNook, Santa Barbara (2015)
32. Robertson, S., Robertson, J.: Mastering the Requirements Process, 3rd edn. Pearson Education, Upper Saddle River (2012)
33. Lucas, J., Bulbul, T., Thabet, W.: An object-oriented model to support healthcare facility information management. Autom. Constr. **31**, 281–291 (2013)
34. Wang, Y., Yu, S., Xu, T.: A user requirement driven framework for collaborative design knowledge management. Adv. Eng. Inform. **33**, 16–28 (2017)
35. Girodon, J., Monticolo, D., Bonjour, E., Perrier, M.: An organizational approach to designing an intelligent knowledge-based system: application to the decision-making process in design projects. Adv. Eng. Inform. **29**, 696–713 (2015)
36. Chong, Y., Chen, C.: Management and forecast of dynamic customer needs: an artificial immune and neural systems approach. Adv. Eng. Inform. **24**, 96–106 (2010)
37. Golzarpoor, B., Haas, C., Rayside, D.: Improving process conformance with industry foundation processes (IFP). Adv. Eng. Inform. **30**, 143–156 (2016)
38. Luo, X., Shen, G., Fan, S.: A case-based reasoning system for using functional performance specification in the briefing process of building projects. Autom. Constr. **19**, 725–733 (2010)
39. Ren, Z., Anumba, C., Augenbroe, G., Hassan, T.: A functional architecture of an e-Engineering hub. Autom. Constr. **17**, 930–939 (2008)
40. Schulz, C., Amor, R., Lobb, B., Guesgen, H.: Qualitative design support for engineering and architecture. Adv. Eng. Inform. **23**, 68–80 (2009)
41. Gadeyne, K., Pinte, G., Berx, K.: Describing the design space of mechanical computational design synthesis problems. Adv. Eng. Inform. **28**, 198–207 (2014)
42. Tucker, S., Bleil de Souza, C.: Thermal simulation outputs: exploring the concept of patterns in design decision making. J. Build. Perform. Simul. **9**(1), 30–49 (2016)

Quality Function Deployment Based Conceptual Framework for Designing Resilient Urban Infrastructure System of Systems

Quan Mao[1], Nan Li[1](✉), and Feniosky Peña-Mora[2]

[1] Department of Construction Management, Tsinghua University, Beijing 100084, China
nanli@tsinghua.edu.cn

[2] Edwin Howard Armstrong Professor of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY 10027, USA

Abstract. The frequent natural hazards and their significant impacts in recent years have repeatedly highlighted the vulnerability of urban infrastructure systems and their lack of system resilience. The ever increasing interdependencies between urban infrastructure systems have further complicated the situation, causing considerable risks of failure propagation. Considering that the properties of urban infrastructure systems including their level of resilience are mainly determined at the design stage, the authors aim to propose a Quality Function Deployment (QFD) based conceptual framework for designing resilient urban infrastructure system of systems (SoS). As a preliminary effort towards this goal, this paper mainly focuses on developing the first matrix in the QFD based framework. Steps to identify resilience criteria and principles are presented, the questionnaire-based method to determine the main body of the first matrix is described, and the approach to work out practical design schemes is explained. Lastly, this paper summarizes the main strength of the proposed framework as well as its limitations, and discusses directions for future research.

Keywords: Infrastructure · System of systems · Resilience criteria Resilience principle · Quality Function Deployment

1 Introduction

Last year was significant in terms of natural hazards. In ten weeks from August to October in 2017, a 124-year-old record was matched with ten consecutive Atlantic storms reaching hurricane strength [1] and causing more than four hundred deaths. Moreover, two devastating earthquakes of magnitude 8.1 and 7.1 hit Mexico City in just two weeks, resulting in over three hundred deaths and massive destruction of buildings [2]. These disasters have highlighted the increasing exposure of urban assets to high-density, large-impact hazards that have caused enormous economic losses reaching hundreds of billions US dollars annually, most of which are in cities with large population [3].

In these situations, urban infrastructure systems (e.g. electric power system, water supply system, transportation system), which provide various services to support fundamental functions of cities, are also becoming quite vulnerable to these hazards. When

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 472–487, 2018.
https://doi.org/10.1007/978-3-319-91635-4_24

studied in all their complexity, urban infrastructure systems can be modeled as a system of systems (SoS) since different infrastructure systems work interactively and interdependently to support the various fundamental functions of urban systems [4]. Within the SoS, the increasing interactions and interdependencies between their different components or facility assets (e.g. substations, water plants) can lead to significant risks of cascading failures. Propagating through these interdependencies, local disaster impacts can grow to become regional ones in a city or global ones in a country. For example, after Hurricane Maria hit Puerto Rico, a complete power outage in the island resulted in the failure of telecommunications and water supplies [5]. In place of electric power, fuel was required for generators to power critical buildings such as hospitals [6]. The competition for fuels made it difficult for trucks to deliver food, water and medicines, which made the situation even worse [6].

These lessons have repeatedly highlighted the lack of resilience in urban infrastructure systems. In this paper, resilience refers to the ability of a system to absorb, adapt to and recover from changes and adverse impacts in the system. Given that urban expansion is expected to continue in the coming decades, building more resilient urban infrastructure systems is a major challenge that urban planners, developers, policy makers and citizens alike are faced with.

The properties of urban infrastructure systems including their level of resilience are mainly determined at the design stage where operation planning and risk management are included [7]. Thus, this paper proposes a conceptual framework for designing resilient urban infrastructure system of systems. It needs to be noted that, given that most cities already exist and will not be designed from scratch, the concept of designing resilience into urban infrastructure systems mostly focuses on new infrastructure development during urban expansion and urban renewal processes, and examines its impact on the resilience level of the entire infrastructure systems including the existing infrastructure components.

2 Related Work

2.1 Approaches for Resilient Urban Infrastructure Systems Design

Various methods or tools have been proposed and developed to support the design of resilient infrastructure systems. Based on traditional reliability and risk assessment techniques, methods used for identifying the weakness of infrastructure systems include failure modes and effects analysis (FMEA) [8], fault and event trees [9] and Bayesian belief networks [10]. These methods can identify and characterize all the risks that could cause a failure, and further identify the failure-critical components. However, the traditional reliability and risk assessment methods regard infrastructure systems as a simple compound of components rather than a SoS. For example, these methods normally regard an electric power system as a network of hardware, with external support such as supervisory control and data acquisition (SCADA), maintenance and emergency management. Such a perspective fails to recognize the electric power system as an integrated and interdependent system composed of not only hardware but various resources and supporting agents, and overlooks the interplays between these system components.

This could be inefficient in resilience design, because most of these methods assume that a system design has existed and suggest using infrastructure components with higher robustness or redundancy to improve the overall resilience [8]. Alternatively, there could be a more efficient design, which achieves improved resilience of the integrated systems by optimizing the interactions between infrastructure systems in a SoS perspective [8].

Another method to design resilient infrastructure systems proposed by Bruneau et al. [11] is to "create" resilience. It measures system resilience based on different properties including robustness, redundancy, resourcefulness and recoverability, and decomposes each single infrastructure system into subsystems on physical, cyber, social and institutional dimensions. Then, the method relates the characteristics (e.g. component capacity) of these subsystems to each property of resilience. Thus, it implies what characteristics can be enhanced to achieve resilience [11, 12]. A significant number of research studies have followed this method and proposed various resilience principles that each single infrastructure system design should meet [12–15]. This method helps transform the complex resilience concept into subsystem characteristics and makes it operable. However, it depends largely on the experience of designers. Without knowing the priority of these characteristics as well as the correlations between them, it is difficult to use this method to support city governments in the allocation of limited resources for building resilience into urban infrastructure systems.

Other methods using high-fidelity simulations have also been proposed for designing resilient infrastructure systems. These methods allow for the investigation of failure propagation and evaluation of different recovery strategies [16]. Each single infrastructure system would be modeled as a system of interactive components by considering their interdependencies. With the resilience assessment metrics such as reduced system performance losses or reduced time of recovery, a trial-and-error process can be conducted to optimize the design scheme by varying infrastructure component characteristics. These methods regard infrastructure systems as a system of systems, and propose an optimized and efficient design scheme for improving resilience. However, it is difficult to set the quantitative mechanism of all interactions between different infrastructure systems with this method since some interdependencies are not well identified or measured (e.g. interdependency between telecommunication system and transportation system). Meanwhile, it is difficult to analyze the correlations that could be synergistic or inhibiting between optimization approaches. This limitation could lead to the impractical design scheme.

2.2 Quality Function Deployment

An approach that has the potential to capture the advantages of all these above methods based on reliability and risk assessment, resilience decomposition and simulation is the QFD. The QFD was developed in Japan in 1966 as a method to transform qualitative customer desires into quantitative engineering parameters that can be controlled [17]. It is an iterative process of transforming customer desires into technical descriptions, successively into component characteristics, process steps and control factors. Thus, it can take into consideration all the stages of development including planning, design, operation and control, and can engage various stakeholders involved in the design and development of complex systems.

House of Quality (HoQ) is the core of QFD as a series of matrices. HoQ can measure the effect of each engineering parameter on each customer desire, or the “relationship” between them, even if they are in separate subsystems. Meanwhile, HoQ can also take the “correlation” that could be synergistic or inhibiting between two engineering parameters into consideration. Based on a specific case study, a self-assessment of customer desires in the existing system can be conducted as well as the expected level of those desires. Based on the gap between these two levels of customer desires, QFD can calculate the specific improvement requirement of given engineering parameters by also taking their implementation difficulty into consideration. This function of QFD method is quite useful to infrastructure systems design.

QFD has been applied in a wide variety of services [18, 19]. For example, it is applied to improve the efficiency and effectiveness of the design of consumer products by transforming customer desires into engineering characteristics and control factors [20, 21]. It is also applied in construction management to support buildable design decision making by analyzing the relationships between client requirements and characteristics of building components [22]. When applied to urban planning, it is used to improve the design of public space by transforming citizen needs into alternative engineering parameters [23].

3 Proposed Conceptual Framework

This section introduces how QFD can be applied into the resilient infrastructure systems design by describing the core concepts, critical steps and main methods in the proposed framework.

3.1 Goals and Core Concepts of the Framework

The main goal of the proposed framework is to decompose the multi-criteria of resilience into a number of engineering parameters, whose relative importance is assessed, during the planning, design, operating and control stages of the lifecycle of infrastructure systems, by using a combination of analysis and simulation methods. The proposed framework also aims to identify factors that have impact on the implementation of engineering parameters, and measure the difficulty of their implementation. Ultimately, the framework can identify the gap between current and expected levels of resilience of the system, and support the decision-making with respect to the investment levels and sequencing of the engineering parameters.

This paper takes the first QoH of QFD (shown in Fig. 1) as an instance to demonstrate the development and implementation process of the proposed framework. The first HoQ mainly focuses on the planning stage of resilient urban infrastructure systems, which concerns the process of transforming customer desires of resilience into engineering parameters considered at the planning stage. The HoQ is composed of a room surrounded by walls on the left and right sides, floor and foundation at the bottom, and ceiling on the top. In addition, there is a triangular roof attached to both the left wall and the ceiling.

Specifically, as shown in Fig. 1, the left "wall" of the HoQ illustrates a list of resilience criteria with their relative importance, while the left "roof" illustrates the correlation between these criteria. Resilience criteria refer to the expected performance of infrastructure systems when responding to extreme events by stakeholders. One examples of resilience criteria is reduced failure consequences of urban infrastructure systems when responding to extreme events. The right "wall" illustrates both the expected and self-assessment levels of each resilience criterion. The "ceiling" illustrates a list of resilience principles, and the top "roof" shows the correlation between these principles. Resilience principles, which are engineering parameters involved at the planning stage, correspond to general system properties such as redundancy and diversity. The "room" illustrates a relationship matrix whose elements reflect the effect of each resilience principle on each resilience criterion. Lastly, the "floor" illustrates the relative importance between resilience principles, and the "foundation" illustrates the implementation difficulty and to what levels each resilience principle should be improved in order to meet the expected resilience level.

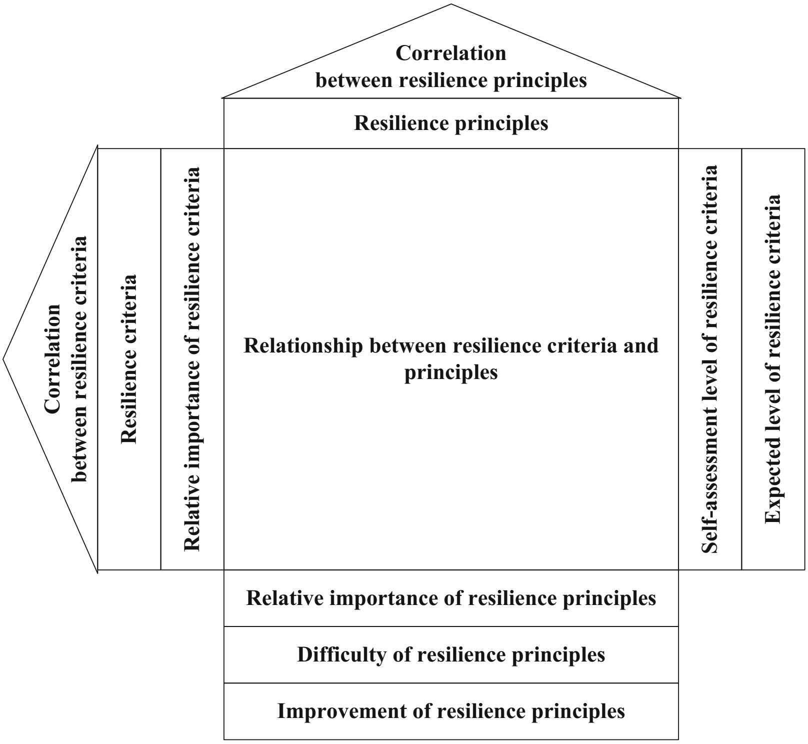

Correlation between resilience principles
Resilience principles
Correlation between resilience criteria
Resilience criteria
Relative importance of resilience criteria
Relationship between resilience criteria and principles
Self-assessment level of resilience criteria
Expected level of resilience criteria
Relative importance of resilience principles
Difficulty of resilience principles
Improvement of resilience principles

Fig. 1. The first HoQ of the proposed framework for designing resilient infrastructure systems

When the relative importance of resilience criteria and the relationship between resilience criteria and principles are provided (e.g. by survey), this HoQ is able to assess the relative importance of resilience principles. Moreover, when the correlation between resilience principles and the implementation difficulty are provided (e.g. by survey), the matrix is also able to assess the level of improvement of each resilience principle that is needed to achieve the expected level of each of the resilience criteria. The following subsections explains in further detail how the proposed framework can be implemented in resilient urban infrastructure systems planning practices.

3.2 Identification of Resilience Criteria and Principles

Urban infrastructure systems are expected to perform to meet certain criteria when responding to extreme events. A review of existing literature was conducted to identify specific infrastructure systems resilience criteria adopted in prior research or practice. According to the definition of each criterion, this paper reduces the number of resilience criteria from six to four, in order to avoid overlap or repetition. For example, failure probabilities and its consequences are so similar that they can vary in the same ways. Meanwhile, the consequences, referring to the decrease of flow or service provided by each single infrastructure system, are more palpable and measurable. Besides, according to the definition, the total performance losses are directly determined by the consequences from failures and recovery time. It could be repetitive to keep them all as separate resilience criteria. Therefore, this paper narrows down the list of resilience criteria into the following: disturbance propagation, consequences from failures, time to recovery, cost to recovery. A list of these criteria with their definitions and sources is presented in Table 1.

Table 1. Resilience criteria, references and examples

Criteria	Definition and references	Examples
Disturbance propagation	Failure propagation due to interdependencies between components of one or several infrastructure systems [24]	Fault-trips propagation in electric power
Consequence from failures	The decrease of flow or service of infrastructure systems [11, 25–34]	Massive black out
Time to recovery	The time from the beginning of disruptive event to full recovery of system functions [11, 24–34]	Time for power system to fully recover from failure
Cost to recovery	The economic cost to restore components and recover system functions [27, 29, 35]	Cost for repairing failed power facilities

Drawing on the application of QFD in other domains, engineering parameters at the planning stage are how a product should be designed as a whole. Based on the existing literature on infrastructure systems resilience design, resilience principles were identified. Resilience principles refer to what resilient infrastructure systems should be designed as. These resilience principles are listed in Table 2 with the source references. Different terms sometimes mean essentially the same principle, and are therefore combined in this paper. For example, adaptability, flexibility, self-regulation, foresight and feedback correction all mean the ability of a system to adapt to changing conditions and undergo a safe failure by changing its configuration. Also, repairability and resourcefulness both mean having adequate resources and personnel to restore the primary failed components directly due to attacks.

Table 2. Resilience principles, references and examples

Principles	Definition and references	Examples
Redundancy	With a number of functionally similar components so that the entire system does not fail when one component fails [8, 11–13, 15, 36–43]	Multiple plants in electric grids; standby pipelines
Diversity	With a number of functionally different components in order to protect the system against various threats [13–15, 36–40, 42–44]	Diverse energy sources; multiple transportation routes
Connectivity	With system components connected so that they support each other [8, 13–15, 36–40, 42–45]	Connected substations; high density of road network
Adaptability	A system should have the ability to "adapt to changing conditions" and undergo a safe failure by changing its configuration [8, 12, 13, 15, 36–42, 44, 45]	Power redistribution responding to disturbances; preparedness based on emergency forecast
Repairability	Ensure availability of adequate resources and personnel to restore the primary failed components directly due to attacks [8, 11–13, 15, 39, 41]	Technical maintenance teams; repairable or replaceable facilities
Independency	A resilient system should possess a "certain degree of self-reliance" that gives it the ability to maintain a minimum acceptable level of functioning (without external support) when influenced by disturbance [13, 15]	Backup power; independent communication channels

3.3 Relative Importance of Resilience Criteria

Following the identification of resilience criteria, it is significant to weight the relative importance between them to support decision-making. The main method used is customer ranking by surveys. In the field of urban planning, the customers should not be limited to end-users but also any stakeholders whose benefits are affected by the outcomes of infrastructure projects and need to be engaged in the decision-making process. Stakeholders involved in resilient infrastructure systems design can be citizens, economic institutions (companies and factories), infrastructure operating institutions and city managers.

The relative importance score between each two resilience criteria is from 1–9 based on Analytic Hierarchy Process (AHP) method [46] (a sample in Fig. 2). The relative importance can be determined by the geometric average of scores from various experts. The number of experts should be more than three and odd [46]. In order to deal with differences of scores between experts, the number of experts is set as three times of score levels, which should be fifty-seven. Finally, based on the judgment matrix, each element of which represents the relative importance between two resilience criteria (see an example in Table 3), the normalized relative importance can be computed by calculating the eigenvector of maximum eigenvalue and normalizing it, respectively based on Eqs. (1) and (2):

$$\left(\lambda_{max}E - R\right)x = 0 \tag{1}$$

$$\bar{x} = \frac{x}{\sum_{1}^{n} x_i} \tag{2}$$

where λ_{max}, x respectively denote the maximum eigenvalue and its eigenvector, E, R denote unit matrix and judgement matrix, $\bar{x}$ denotes the normalized importance weights, and x_i denotes the i-th element of vector x. In the judgment matrix, there could be a contradictory situation where relative importance between several items is not consistent. In such case, an approach termed consistency test is used to ensure the reliability of the results [46]. It estimates the consistency level of different relative importance in the judgment matrix by comparing maximum eigenvalue with matrix dimensions.

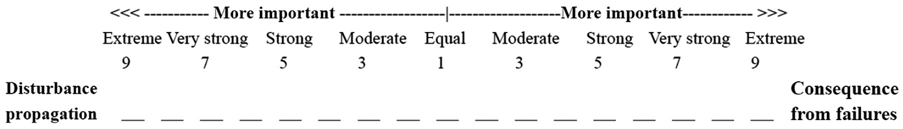

<<< ---------- More important -----------------|------------------More important----------- >>>
Extreme Very strong Strong Moderate Equal Moderate Strong Very strong Extreme
9 7 5 3 1 3 5 7 9
Disturbance propagation
Consequence from failures

Fig. 2. Example of relative importance measurement of resilience criteria

Table 3. Example of judgment matrix of resilience criteria

	Disturbance propagation	Consequence from failures	Cost to recovery	Time to recovery
Disturbance propagation	1	1/3	5	1/2
Consequence from failures	3	1	7	1/3
Cost to recovery	1/5	1/7	1	1/9
Time to recovery	2	3	9	1

3.4 Relationship Between Resilience Criteria and Principles

The main strength of QFD is to transform the customer desires into engineering parameters that can be controlled based on the relationship between them. In the proposed framework, each resilience principle can have an effect on certain resilience criteria. For example, redundancy (resilience principle) could reduce the consequences from failure (resilience criterion). The relationship matrix reflects the strength of such effect. The strength can be weighted on a scale of: 9 (extremely strong), 5 (very strong), 1 (weak), and 0 (no relationship) following works in other applications of QFD [17, 47]. The identification of the relationship and its specific score can be determined by several methods [48]:

- Brainstorming based on technical knowledge;
- Expert scoring including different stakeholders;
- Design experiments;
- Historical product data analysis.

The method of simulation or experiments is difficult to implement due to the large-scale and complexity of urban infrastructure systems. This paper proposes three approaches to reveal the relationship between resilience criteria and principles. Firstly, the relationship is identified based on literature review and brainstorming based on technical knowledge. The results (shown in Table 4) are qualitative and can be used to validate the results of the latter two methods.

The second approach is based on the direct score of the relationship by experts in a survey. Given the explanation of each principle and criterion, experts are asked to provide an assessment (9-strong, 5-moderate, 1-weak or 0-no) to estimate the effect of each principle on each criterion based on their experience in the infrastructure domain. The reliability of results is ensured by consistence test using Kappa coefficient [49] and the validity can be ensured by comparing the results from different stakeholders.

The third approach, which is also survey-based, uses questions developed based on actual cases and data to decrease the dependence of results on experts' experiences and avoid subjectivity. The survey lists a number of typical existing infrastructure investment/design/construction projects. Experts are asked to select those projects that they ever participated in or are familiar with, and estimate the implementation level (1–5 scale) of each resilience principle in these selected projects (e.g. they are asked "how much redundancy do you think the single infrastructure system has in this project"). They are also asked to estimate the extent to which each resilience criterion is achieved in these projects (e.g. they are asked "how severe do you think the consequences from failures in this project were during past hazard events or would be during a virtual hazard event"). To decrease the subjectivity of these assessments, all survey respondents are presented with the same descriptions of these projects and hazard events. To yield statistically reliable results, the number of experts should be five to ten times the number of questions [50]. Given that there are four resilience criteria and six resilience principles to be assessed in the survey, a total of 50–100 responses are needed for each project included in the survey. Then a structure equation model (SEM) can be built to analyze the survey data. SEM is a statistical framework for analyzing the relationships among a collection of variables simultaneously in one model with a diverse set of mathematical models, computing algorithms, and statistical methods [51]. Exploratory factor analysis can be conducted to test the validity of each item in resilience criteria or principles. Meanwhile, path analysis can be conducted to reveal the relationship between each resilience criterion and principle in a quantitative way. To be used in QFD, the strength of relationship should then be transformed to a scale of 1–9.

Table 4. Identification of relationship between resilience criteria and principles

	Disturbance propagation	Consequence from failures	Time to recovery	Cost to recovery
Redundancy	Redundancy can provide backup goods or service to avoid the disturbance propagation caused by lacking power or service	Redundancy can provide backup goods or service to mitigate the impact of failures	No relationship	Fewer failures mean less cost
Diversity	Diversity can provide backup goods or service to avoid the disturbance propagation caused by lacking power or service	Diversity can provide backup goods or service to mitigate the impact of failures	Diversity can provide diverse ways to make the restoration work more efficient	Fewer failures mean less cost
Connectivity	Connectivity can aggravate the situation due to the interaction	Connectivity can aggravate the disturbance propagation due to the interaction	Connectivity can provide diverse ways to make the restoration work more efficient	Fewer failures mean less cost
Adaptability	Adaptability can mitigate the disturbance propagation by adjustments and self-regulation	Adaptability can mitigate the consequences by adjustments and self-regulation	Adaptability can accelerate the restoration work by adjustments and self-regulation	Fewer failures mean less cost
Repairability	No relationship	No relationship	Repairability can reduce the restoration time with adequate resources	
Independency	Independency can make subsystem not influenced by external failures	No relationship	No relationship	Fewer failures mean less cost

3.5 Relative Importance of Resilience Principles

Based on the function of QFD, the relative importance of resilience principles is worked out with the relative importance of resilience criteria and the relationship between resilience criteria and principles, which can be described as Eq. (3):

$$p_j = \sum\nolimits_{i=1}^{4} c_i A_{ij}. \tag{3}$$

where c_i and p_j denote respectively the relative importance of resilience criterion i and principle j, and A_{ij} denotes the relationship between resilience criterion i and principle j.

3.6 Correlation Between Resilience Criteria or Resilience Principles

For an integrated infrastructure SoS, engineering parameters are inter-correlated, and it is difficult to change one without affecting the others. For example, connectivity and independency cannot be enhanced simultaneously (e.g. increasing connectivity of nodes in a water supplies system always leads higher dependency), while diversity and adaptability can be enhanced at the same time (e.g. with diverse energy sources, the demand of power can be satisfied by adjusting supplies of different sources when some of them are lacking). This paper proposes two survey-based approaches to identify such correlations. The first approach is based on the direct score of the correlation by experts in a survey. The score has a scale of: −2 for strong negative correlation, −1 for negative correlation, 0 for no correlation, 1 for positive correlation, 2 for strong positive correlation. The second approach is based on the survey of resilience assessment in listed projects. The correlation between resilience criteria and principles can be analyzed with the SEM. To be used in QFD, the correlation should then be transformed to a scale of −2–2.

3.7 Difficulty of Implementation of Resilience Criteria and Principles

Due to limited resources or technology, the implementation priority of resilience criteria or principles should be determined based on not only their relative importance but also their difficulty of implementation. For example, redundancy is usually preferred to achieve resilience but requires significant resources. Hence, it is significant to estimate the difficulty of implementation of resilience criteria and principles in practice. To this end, this paper proposed a survey-based approach. Experts are asked to provide a score of 1–5 (1-extreme easy and 5-extrem difficult) to estimate the implementation difficulty of each resilience criterion or principle based on their experience in the infrastructure domain. Possible factors that may impact the level of implementation difficulty include but are not limited to time, budget, and government policy. Moreover, this paper identifies several possible factors including time pressure, political pressure, stakeholder pressure and community based on literature review, which will be expanded by expert interview in future work. Experts are also asked to provide a score of 1–5 (1-weak and 5-strong) to estimate the effect of each factor on implementing resilience based on their experience in the infrastructure domain.

3.8 Improvement of Resilience Principles

The main purpose of this framework is to provide practical guidance of resilient infrastructure systems design. After resilience criteria are transformed into actionable resilience principles and the relative importance of resilience principles are assessed, the next step is to determine which resilience principles should be prioritized in infrastructure systems investment decision making, so as to maximize the resilience level of the infrastructure systems given the constraints of availability of resources. Based on the method of QFD, the gap (Δy) between the real and expected level of resilience criteria should be investigated first. Using the relationship (A) between resilience criteria and principles, the gap can be transformed into the improvement requirement ($\varDelta x$) of resilience principles as shown in Eq. (4):

$$A\varDelta x^{T} = \Delta y \tag{4}$$

Taking correlation between resilience principles into consideration, an optimized solution of improvement of resilience principles can be worked out to promote principles that have positive correlation and avoid those that have negative correlation. The correlations are shown as Eq. (5):

$$\mathrm{a}\Delta x_i \pm b\Delta x_j + c_{ij} = 0 \; i,j = 1 \sim n \tag{5}$$

where Δx_i and Δx_j denote the change of i-th and j-th resilience principles, respectively, c_{ij} denotes the correlation strength, n denotes the number of principles, and a, b denote constant parameters. Moreover, to achieve minimal implementation difficulty (d), with the constraints described in Eqs. (4) and (5), a more practical solution to the improvement of resilience principles can be worked out based on Eq. (6):

$$min\left(d\varDelta x^{T}\right) \tag{6}$$

4 Outlook and Limitations

Based on the above explanation of implementation details of the proposed framework, the main strength of the framework can be summarized as follows:

- it could transform qualitative resilience criteria into engineering parameters (e.g. resilience principles) that can be addressed in practice;
- it could take into consideration all the stages (e.g. planning, design, operation, control) and involve different stakeholders (e.g. government employee, city planner, emergency personnel, public safety officer, architect, engineer, contractor, infrastructure operator, infrastructure investor) taking into consideration the complexity of resilient infrastructure systems design;
- it is applicable to a complex system since it could regard the system as a system of systems, by using the correlation of items in row and column;

- it could work out the specific improvement requirement of given items based on the gap between expected and real resilience criteria;
- it could optimize the solution of improvement of resilience principles by taking correlation between resilience principles and implementation difficulty into consideration.

There are also several limitations of the proposed framework that should be addressed in future research:

- it mainly depends on the literature review and survey and the scoring is subjective;
- the reliability of results may be hard to be ensured since different stakeholders have different understanding of the listed items;
- the respondents may be impatient with the lengthy questionnaire that includes a number of sections including relative importance, relationship, correlation, implementation difficulty.
- the implementation cost of the questionnaire could be higher compared to simulation since as it involves multiple stakeholders and a good number of experts.

5 Conclusions

This paper proposed a conceptual framework for designing resilient urban infrastructure system of systems based on QFD. It takes the first HoQ of QFD as an example to explain the development and implementation process of this framework. The first HoQ mainly focuses on the planning stage of urban infrastructure systems. The identification of resilience criteria and resilience principles based on literature review were explained firstly. Then the paper elaborated on a survey-based approach used to construct the main body of the HoQ. Different components of the HoQ represented the relative importance between resilience criteria, relationship between resilience criteria and principles, correlation between resilience criteria or principles, and implementation difficulty of resilience principles. Lastly, the main functions of QFD were introduced which could be used to optimize practical design schemes. Future research will be carried out by the authors to materialize the first HoQ with surveys and associated analysis, develop following HoQ that are related to later phases of the lifecycle of infrastructure systems, and integrate fuzzy decision-making and simulation as alternative approaches to the interpretation of survey results and construction of the HoQ.

Acknowledgments. This material is based upon work supported by National Key R&D Program of China under grant No. 2017YFC0803308, National Natural Science Foundation of China (NSFC) under grant No. U1709212 and 71741023, and Tsinghua University Initiative Scientific Research Program under grant No. 2014z21050 and 2015THZ0. The authors are thankful for the support of Ministry of Science and Technology of China, NSFC and Tsinghua University. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. New York Times. https://mobile.nytimes.com/2017/10/11/climate/hurricane-ophelia.html?emc=edit_th_20171012&nl=todaysheadlines&nlid=60043584&referer=. Accessed 20 Dec 2017
2. Cable News Network (CNN). http://edition.cnn.com/2017/09/19/americas/mexico-earthquake/index.html. Accessed 20 Dec 2017
3. UNISDR. http://www.preventionweb.net/english/hyogo/gar/2015/en/gar-pdf/GAR2015_EN.pdf. Accessed 22 Dec 2017
4. Kasai, S., Li, N., Fang, D.: A system-of-systems approach to understanding urbanization - state of the art and prospect. Smart Sustain. Built Environ. **4**, 154–171 (2015)
5. Telegraph News. http://www.telegraph.co.uk/news/2017/09/20/hurricane-maria-path-latest-news-live-puerto-rico-virgin-islands/. Accessed 25 Dec 2017
6. Puerto Rico Economic Recovery Initiative. http://fixpuertorico.org/2017/09/29/cnbc-puerto-rico-short-on-fuel-cannot-deliver-food-and-medicine-to-the-victims-of-hurricane-maria/. Accessed 25 Dec 2017
7. Childers, D.L., Cadenasso, M.L., Grove, J.M., Marshall, V., McGrath, B., Pickett, S.T.A.: An ecology for cities: a transformational nexus of design and ecology to advance climate change resilience and urban sustainability. Sustainability **7**(4), 3774–3791 (2015)
8. Uday, P., Marais, K.: Designing resilient systems-of-systems: a survey of metrics, methods, and challenges. Syst. Eng. **18**(5), 491–510 (2015)
9. Fleming, C.H., Spencer, M., Thomas, J., Leveson, N., Wilkinson, C.: Safety assurance in NextGen and complex transportation systems. Saf. Sci. **55**, 173–187 (2013)
10. Aven, T.: On how to deal with deep uncertainties in a risk assessment and management context. Risk Anal. **33**, 2082–2091 (2013)
11. Bruneau, M., Chang, S., Eguchi, R., Lee, G., O'Rourke, T., Reinhorn, A., Shinozuka, M., Tierney, K., Wallace, W., von Winterfeldt, D.: A framework to quantitatively assess and enhance the seismic resilience of communities. Earthq. Spectra **19**, 733–752 (2003)
12. Jackson, S., Ferris, T.L.J.: Resilience principles for engineered systems. Syst. Eng. **16**, 152–164 (2013)
13. Sharifi, A., Yamagata, Y.: Major principles and criteria for development of an urban resilience assessment index. In: Proceedings of 2014 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE). IEEE, New York (2014)
14. Farid, A.M.: Multi-agent system design principles for resilient operation of future power systems. In: Proceedings of 2014 IEEE International Workshop on Intelligent Energy Systems (IWIES), pp. 18–25. IEEE, New York (2014)
15. Sharifi, A., Yamagata, Y.: Principles and criteria for assessing urban energy resilience: a literature review. Renew. Sustain. Energy Rev. **60**, 1654–1677 (2016)
16. Alessandri, A., Filippini, R.: Evaluation of resilience of interconnected systems based on stability analysis. In: Hämmerli, Bernhard M., Kalstad Svendsen, N., Lopez, J. (eds.) CRITIS 2012. LNCS, vol. 7722, pp. 180–190. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41485-5_16
17. Akao, Y.: Quality Function Deployment (QFD) Integrating Customer Requirements into Product Design. Productivity Press, Porland (1990)
18. Chen, L., Ko, W., Tseng, C.: Fuzzy approaches for constructing house of quality in QFD and its applications: a group decision-making method. IEEE Transp. Eng. Manag. **60**, 77–87 (2013)

19. Kuo, M., Lee, T.: Application of house of quality (HOQ) to health care management. In: McDaniel, J. (ed.) Conference on Revolutionizing Health Care with Informatics 2009, 143, pp. 371–375. IOS Press, Netherlands (2009)
20. Li, Y., Huang, M., Chin, K., Luo, X., Han, Y.: Integrating preference analysis and balanced scorecard to product planning house of quality. Comput. Ind. Eng. **60**, 256–268 (2011)
21. Belhe, U., Kusiak, A.: The house of quality in a design process. Int. J. Prod. Res. **34**, 2119–2131 (1996)
22. Yang, Y., Wang, S., Dulaimi, M., Low, S.: A fuzzy quality function deployment system for buildable design decision-makings. Autom. Constr. **12**, 381–393 (2003)
23. Wey, W., Chiu, Y.: Assessing the walkability of pedestrian environment under the transit-oriented development. Habitat Int. **38**, 106–118 (2013)
24. Reed, D.A., Kapur, K.C., Christie, R.D.: Methodology for assessing the resilience of networked infrastructure. IEEE Syst. J. **3**(2), 174–180 (2009)
25. Yodo, N., Wang, P.: Engineering resilience quantification and system design implications: a literature survey. J. Mech. Des. **138**(11), 111408 (2016)
26. Francis, R., Bekera, B.: A metric and frameworks for resilience analysis of engineered and infrastructure systems. Reliab. Eng. Syst. Saf. **121**, 90–103 (2014)
27. Ouyang, M., Duenas-Osorio, L., Min, X.: A three-stage resilience analysis framework for urban infrastructure systems. Struct. Saf. **36–37**, 23–31 (2012)
28. Cimellaro, G.P., Reinhorn, A.M., Bruneau, M.: Framework for analytical quantification of disaster resilience. Eng. Struct. **32**, 3639–3649 (2010)
29. Henry, D., Emmanuel, R.-M.J.: Generic metrics and quantitative approaches for system resilience as a function of time. Reliab. Eng. Syst. Saf. **99**, 114–122 (2012)
30. Chang, S., Shinozuka, M.: Measuring improvements in the disaster resilience of communities. Earthq. Spectra **20**, 739–755 (2004)
31. Representing perceived tradeoffs in defining disaster resilience: 30. Zobel C. W. Decis. Support Syst. **50**, 394–403 (2011)
32. Barker, K., Ramirez-Marquez, J., Rocco, C.M.: Resilience-based network component importance measures. Reliab. Eng. Syst. Saf. **117**, 89–97 (2013)
33. Ouyang, M., Duenas-Osorio, L.: Time-dependent resilience assessment and improvement of urban infrastructure systems. Chaos **22**(3), 033122 (2012)
34. Bruneau, M., Reinhorn, A.: Exploring the concept of seismic resilience for acute care facilities. Earthq. Spectra **23**, 41–62 (2007)
35. Vugrin, E.D., Warren, D.E., Ehlen, M.A.: A resilience assessment framework for infrastructure and economic systems: quantitative and qualitative resilience analysis of petrochemical supply chains to a hurricane. Process Saf. Prog. **30**(3), 280–290 (2011)
36. Godschalk, D.R.: Urban hazard mitigation: creating resilient cities. Natural Hazards Rev. **4**(3), 136–143 (2003)
37. Biggs, R., Schluter, M., Biggs, D., Bohensky, E.L., Burnsilver, S., Cundill, G., Dakos, V., Daw, T.M., Evans, L.S., Kotschy, K., Leitch, A.M., Meek, C., Quinlan, A., Raudsepp-Hearne, C., Robards, M.D., Schoon, M.L., Schultz, L., West, P.C.: Toward principles for enhancing the resilience of ecosystem services. In: Gadgil, A., Liverman, D. (eds.) Annual Review of Environment and Resources, 37, pp. 421–448. Annual Reviews, Palo Alto (2012)
38. Wiese, F.: Resilience thinking as an interdisciplinary guiding principle for energy system transitions. Resources-Basel **5**(4), 30 (2016)
39. Bruijn, K., Buurman, J., Mens, M., Dahm, R., Klijn, F.: Resilience in practice: five principles to enable societies to cope with extreme weather events. Environ. Sci. Policy **70**, 21–30 (2017)

40. Clarvis, M.H., Bohensky, E., Yarime, M.: Can resilience thinking inform resilience investments? Learning from resilience principles for disaster risk reduction. Sustainability **7**(7), 9048–9066 (2015)
41. Liao, K., Le, T.A., Van Nguyen, K.: Urban design principles for flood resilience: Learning from the ecological wisdom of living with floods in the Vietnamese Mekong Delta. In: International Symposium on Ecological Wisdom for Urban Sustainability, vol. 155, pp. 69–78. Elsevier Science BV, Netherlands (2016)
42. Mauthe A., Hutchison D., Cetinkaya E.K., Ganchev I., Rak J., Sterbenz J.P.G., Gunkel M., Smith P., Gomes T.: Disaster-resilient communication networks: principles and best practices. In: Jonsson, M., Rak, J., Somani, A., Papadimitriou, D., Vinel, A. (eds.) Proceedings of 8th International Workshop on Resilient Networks Design and Modeling (RNDM), pp. 1–10. IEEE, New York (2016)
43. Sterbenz, J.P.G., Hutchison, D., Cetinkaya, E.K., Jabbar, A., Rohrer, J.P., Schoeller, M., Smith, P.: Resilience and survivability in communication networks: strategies, principles, and survey of disciplines. Comput. Netw. **54**(8), 1245–1265 (2010)
44. Anderies, J.M.: Embedding built environments in social-ecological systems: resilience-based design principles. Build. Res. Inf. **42**(2), 130–142 (2014)
45. Clarvis, M.H., Allan, A., Hannah, D.M.: Water, resilience and the law: from general concepts and governance design principles to actionable mechanisms. Environ. Sci. Policy **43**, 98–110 (2014)
46. Saaty, T.L.: The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. RWS Publications, Pittsburgh (1988)
47. Khoo, L.P., Ho, N.C.: Framework of a fuzzy quality function deployment system. Int. J. Prod. Res. **34**(2), 299–311 (1996)
48. Verma, R., Maher, T., Pullman, M.: Effective product and process development using quality function deployment. In: Usher, J.M., Roy, U., Parsael, H.R. (eds.) Integrated Product and Process Development, pp. 339–354. Wiley, New York (1998)
49. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**, 37–46 (1960)
50. Krosnick, J.A., Presser, S.: Question and questionnaire design. In: Marsden, P.V., Wright, J.D. (eds.) Handbook of Survey Research, pp. 263–313. Emerald, England (2010)
51. Hoyle, R.H.: Structural Equation Modeling: Concepts, Issues, and Applications. Sage Publications, Thousand Oaks (1995)

Quantifying Performance Degradation of HVAC Systems for Proactive Maintenance Using a Data-Driven Approach

Gokmen Dedemen and Semiha Ergan(✉)

New York University, Brooklyn, NY 11201, USA
semiha@nyu.edu

Abstract. Poorly maintained and degraded Heating, Ventilating, Air Conditioning (HVAC) systems waste significant amount of energy. Current Facilities Management (FM) practice is mostly based on reactive and scheduled maintenance of HVAC systems instead of proactive maintenance, which aims at detecting anticipated failures before they occur, so that lower life cycle costs can be accomplished. Therefore, current FM practice needs approaches to detect anticipated failures, so that proactive measures can be taken. Building Automation Systems (BASs) in smart buildings provide historical data on HVAC operations, which can be leveraged for detecting performance degradation of HVAC systems. This study provides a data-driven methodology to quantify and visualize performance changes of HVAC systems over the years using historical BAS data. Our results on a case building demonstrated that there are statistically significant differences between the dataset over the years due to behavioral changes in the HVAC system when other factors (e.g., weather) are controlled. The contribution of this work is a computational approach to identify behavioral changes in HVAC equipment over time using custom selected algorithms for the HVAC domain.

Keywords: HVAC · Proactive maintenance · Data driven approaches

1 Introduction

It has been reported that poorly maintained and degraded Heating, Ventilating, Air Conditioning (HVAC) systems waste around 15% to 30% of energy used in commercial buildings [1]. Since buildings consume 40% of the total energy consumption in the U.S. [2], improper maintenance and degradation of HVAC systems can be perceived as a significant source of energy waste and a major obstacle in the way of attaining sustainable and energy efficient buildings.

A limited number of approaches has been utilized in the current Facility Management (FM) practice for effective maintenance of HVAC systems, such as reactive and scheduled maintenance. However, they are either based on replacing or repairing the equipment after failure occurs (reactive maintenance), or performing maintenance tasks periodically, such as cleaning filters once in a year (scheduled maintenance) [3]. These approaches are not considered as effective ways of maintaining HVAC systems, since they do not offer ways to detect failures before they occur, leading to high

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 488–497, 2018.
https://doi.org/10.1007/978-3-319-91635-4_25

equipment life cycle costs. Therefore, current FM practice needs approaches to detect anticipated failures in a system, so that proactive measures can be taken promptly, and low life cycle costs can be accomplished.

Intelligent buildings in the modern society provide vast amount of data on actual HVAC operations through Building Automation Systems (BASs) [4]. Therefore, BASs can serve as a historical data source in order to detect anticipated failures, or performance degradation in HVAC systems. We envision to use such historical performance data to detect patterns of change in system behaviors. The objective of this study is to develop a data-driven approach to quantify performance degradation of HVAC systems over the years using BASs data. For this purpose, the authors proposed a four-step data-driven approach to quantify performance degradation of Air Handling Units (AHUs) of HVAC systems over the years. The approach was tested using the data measured on one of the air handling units (AHU) of a case building. The building analyzed in this study is a three-story highly sensed building, where AHU parameters and its electricity consumption have been collected for every minute. The available BAS data involves AHU operational parameters, such as supply air temperature, humidity, and fan speed. After data pre-processing, the first step of the proposed approach is to apply motif discovery techniques on the electricity consumption data of AHUs in order to identify similar daily patterns that occur in different years. In the second step, feature selection methods are performed in order to identify the operational parameters of the AHUs analyzed influencing AHU electricity consumption, so that a multi-dimensional dataset (such as supply air temperature, flowrate, and fan speed) is obtained. This multi-dimensional dataset contains the operational parameters dictating the energy performance of the system. In the final step, the multi-dimensional dataset obtained from feature selection process for each year analyzed has been visualized using Multi-Dimensional Scaling (MDS). Since the data of three consecutive years was analyzed in our study, for a given motif, MDS produced three data points, each data point representing yearly data. The contribution of this paper is a computational approach to leverage BAS data to understand changes in operational behavior of HVAC equipment over the years. The computing contribution of this work is a four-step computational approach to detect such changes in system behaviors. The paper formalizes a step by step approach to implement a custom set of data driven algorithms to analyze any BAS data for detecting operational performance changes.

2 Background and Customization of Algorithms

With the installation of BASs in today's buildings, the data on HVAC operations has become easily accessible. Therefore, data-driven fault detection studies in HVAC systems are in abundance. Data-driven fault detection studies intend to detect faults that occur in HVAC systems by using readily available BASs data. Data-driven fault detection studies provide numerous advantages over other types of fault detection methods (i.e., Model based and Rule based), since they do not require any physical model of HVAC systems, or domain knowledge in advance [5, 6]. On the other hand, they usually rely on supervised learning, which requires a training data set with labeled instances as faulty and not-faulty [7]. Moreover, sufficient number of faulty instances should exist in the dataset in order to successfully train the supervised algorithms. These limits the applicability of supervised-learning based approaches on HVAC systems. On the other hand, unsupervised approaches do not require a training process where abnormal instances are labeled, instead they are based on pattern recognition methods, where regularities and irregularities are detected in the dataset. The existence of irregularities is associated with the performance degradation, or system faults. Therefore, our study utilizes an unsupervised approach in order to quantify and visualize performance degradation of a system.

Various unsupervised approaches have been proposed in the literature in order to detect faults in HVAC systems, such as clustering, Principal Component Analysis, and Bayesian Networks [8]. However, they aim to find regularities (patterns) in the data independent of time, whereas BAS data has a temporal dimension. Therefore, a temporal data-driven approach is required to address degradation problem. We utilized Symbolic Aggregate Approximation (SAX), which is a temporal data-driven approach, for pattern recognition. It transforms a time series data into symbolic representations as strings for computationally efficient processing of datasets [9]. Details are provided later in the implementation section.

Another obstacle when dealing with the BAS data is the large number of variables involved in the dataset. BAS dataset contains variables that do not depict electricity usage information. Therefore, there is a need for filtering, or reducing the number of variables in the BAS dataset, such that it will only contain AHU operational parameters that depict the electricity use information. In other words, there is a need for a feature, or variable selection process to identify relevant information. In this study, wrapper feature selection algorithms have been used since they provide more accurate results compared to other types of feature selection methods (e.g., filter and embedded algorithms) [10].

Although feature selection methods reduce the dimensionality in the dataset (i.e., the number of variables), visualization of the BASs data is still a challenge, if the dataset involves more than three variables. This is because only three-dimensional data (x, y, and z) can be visualized. Therefore, a dimensionality reduction method is needed to effectively visualize available data. An effective dimensionality reduction method requires preserving the relative inter-point distances, so that the characteristics of the data can be preserved, while visualizing the data. Therefore, Multi-Dimensional Scaling (MDS) has been selected to be used in this study, since it preserves the relative distances between data points after transforming the dataset into lower dimensions (usually two dimensions) [11]. In this study, MDS was used to visually represent the variables after feature selection process using a Cartesian coordinate system (two dimensions).

It is clear from the literature that there is a need for defining behavioral changes in HVAC equipment with time considered as an influential factor. The study presented in this paper differs from previously published work on the same domain by developing a computational approach that considers temporal dimension in the sensor data being analyzed. The study also differs from previous research studies on the developed data-driven approach custom to the BAS data analysis.

3 Research Method and Implementation

The proposed methodology consists of the data-preprocessing, motif discovery, feature selection, and dimensionality reduction steps to process BASs data. In this regard, the methodology requires raw BAS data as input.

3.1 Overview of the Utilized Dataset

Minute interval data for consecutive three years from an air handling unit (AHU) of an existing building along with electricity consumption of that AHU has been utilized to evaluate the proposed methodology. The data was captured for every 24-h period and every day in a week continuously. Available data set involves measurements of forty AHU operational parameters besides AHU energy consumption. The operational parameters included physical measurements through the system such as supply air temperature, return air pressure, and fan speed.

3.2 Step 1: Data Pre-processing

Data pre-processing aims to increase the quality of the data for improving the performance of the data-driven algorithms applied on a given dataset [12]. In this study, if a missing value was detected for a given timestamp for any of the parameters, the whole data recorded at that timestamp for all the remaining parameters was eliminated from the dataset. An outlier was considered as the data beyond three standard deviations from its mean, and removed from the dataset, as practices by similar studies using BAS data [13]. In addition, data normalization was applied in order to ensure that the analyzed AHU operational parameters did not have varying scales. Data normalization in this study relied on z-normalization, which reshaped the parameters analyzed such that each parameter has a mean of 0, and standard deviation of 1.

3.3 Step 2: Motif Discovery

SAX necessitates two inputs to be defined in advance: a window size (w) and an alphabet size (a). A normalized time series data, with a size of $n,$ is divided into w windows such that the number of data points in each window is equal. Then, based on the selection of the alphabet size, each window is assigned a letter. These letters possess qualitative information, such as low (letter a), medium (letter b), and high (letter c) [9]. For SAX, the window size (w) was selected as 6 h, and the alphabet size (a) was selected as three as suggested in the similar studies using SAX on BAS data [13]. Therefore, a daily electricity usage pattern was represented by a string composed of 4 letters each representing the 6-h time period in a day.

After the data pre-processing step on AHU operational parameters, AHU electricity consumption data was included for 365 days for the first two years, and 317 days for the last year (due to data unavailability in the last year). SAX was applied on AHU electricity use data to identify daily patterns in the form of four letter strings. Daily electricity use patterns identified by SAX algorithm in each year are provided in Table 1. Most frequently occurring patterns (i.e., motifs) that are common in three consecutive years were identified, and highlighted as shown in Table 1. Table 1 shows the whole number of days and corresponding patterns. Four common motifs were identified as “*bcca*”, “*aaaa*”, “*bbbb*”, and “*cccc*”, where *a* represents ‘low consumption’, *b* represents ‘medium consumption’, and *c* represents ‘high consumption’. The criteria used to identify common motifs was that a motif should be common in all three years, and its frequency should not be less than 10 days. While the motif “*aaaa”* represents AHU electricity consumption in the weekends, where the electricity consumption is minimum (represented by *a* in each 6-h interval), the other three patterns represent weekday motifs.

Table 1. SAX results in Years 1, 2, and 3.

Year 1		Year 2		Year 3	
Pattern	Frequency	Pattern	Frequency	Pattern	Frequency
bbca	125	bbbb	89	cccc	86
aaaa	72	bcca	69	bcca	43
bccb	32	bccb	41	aaaa	43
bbbb	22	bbba	34	abba	36
bbba	20	aaaa	22	bbbb	35
cccc	19	ccca	20	bbba	18
bcba	19	cccc	16	ccca	9
bbca	14	cccb	15	bccb	8
bbcb	8	abbb	13	abbb	7
bccc	6	aaba	8	aabb	5
abbb	4	abba	7	acca	5
aaac	3	aabb	7	bccc	3
cccb	3	bccc	5	accb	3
ccca	3		3	abca	2
aacc	2	bcba	2	bbbc	2
bcbb	2	abcb	2	bbcc	2
ccba	2	aaab	2	cccb	2
acca	2	bbcb	2	aaab	1
aaab	1	cbbb	1	bbab	1
aabc	1	bcbb	1	aaba	1
aaba	1	abcc	1	aaba	1
bbcc	1	abca	1	ccbc	1
abba	1	abaa	1	abcc	1
accc	1	bbca	1	acba	1
ccbb	1	cbcc	1	bbca	1
		accb	1		

3.4 Step 3: Feature Selection

As mentioned in the background section, wrapper feature selection algorithms were selected to be used in this study. Among wrapper feature selection methods, Recursive Feature Elimination with Cross Validation (RFECV) was used in this study because of its wide use and higher accuracy shown by previous studies [10].

After the motif discovery, feature selection step was performed by using RFECV. The output of the RFECV algorithm provides the subset of variables that dictate the level of AHU electricity usage. AHU performance parameters identified as drivers of the AHU electricity consumption are shown in Table 2.

Table 2. Feature selection results.

Parameter	Description	Unit
SAT	Supply air temperature	F
RAP	Return air pressure	psi
SAP	Supply air pressure	psi
OAP	Outside air pressure	psi
PAH	Pressure after heating	psi
PAC	Pressure after cooling	psi
FSPD	Fan speed	rpm
OAV	Outside air velocity	fpm
FAF	Fresh air flow	cfm
SAF	Supply air flow	cfm

3.5 Step 4: Dimensionality Reduction

Dimensionality reduction techniques transform multi-dimensional data into lower dimensions. In this study, Multi-Dimensional Scaling (MDS) has been used to overcome dimensionality barrier, since it preserves the distances between data points, hence preserving the characteristics of the data.

In the final step, the multi-dimensional dataset obtained from feature selection process, in other words, the dataset involving 10 variables shown in Table 2 has been visualized using Multi-Dimensional Scaling (MDS). As mentioned in the background section, MDS enables visualization of multi-dimensional data on Cartesian coordinate system (2 dimensions) to understand the level of similarity of individual cases in datasets. In this study, MDS was used to visualize the dataset involving 10 variables shown in Table 2 corresponding to the AHU electricity usage motifs identified in Table 1 (‘*aaaa*’, ‘*bbbb*’, ‘*bcca*’, and ‘*cccc*’). The outputs of MDS are data points in Cartesian coordinate system for each motif and for each year (Figs. 1 and 2). Since the data of three consecutive years was analyzed in our study, for a given motif, MDS produced three data points, each data point representing yearly data representing the 10 variables that are shown in Table 2.

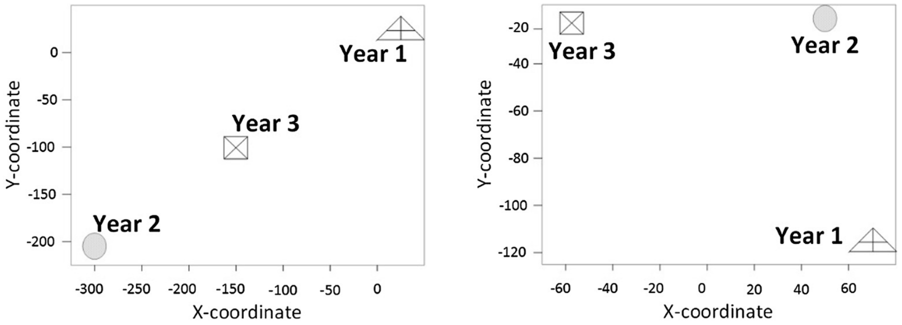


Fig. 1. MDS visualization for the energy consumption motifs "*aaaa*" (left) and "*bbbb*" (right). Triangle: Year 1; Circle: Year 2; Square: Year 3.

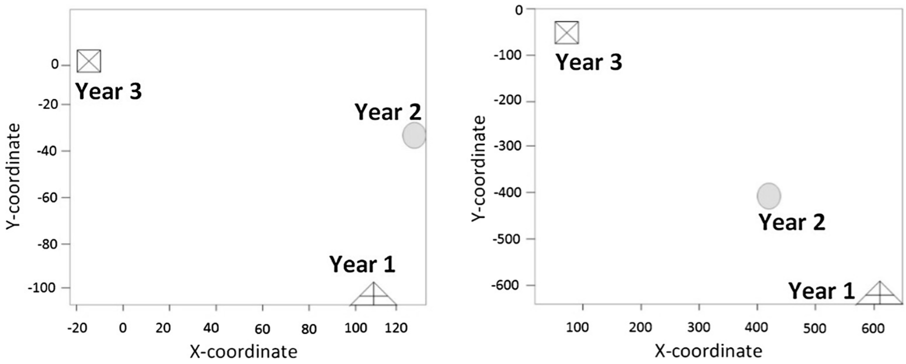


Fig. 2. MDS visualization for the energy consumption motifs "*bcca*" (left) and "*cccc*" (right).

Since MDS preserves inter-point distances between the data analyzed, if the points given by MDS are in close proximity with each other, then, the difference between dataset of different years are close to each other. On the other hand, if the points are far away from each other, it signals that there are significant differences between data analyzed for different years. Ideally, it would be expected that for a given pattern/motif, the points for the given years would be very close to each other- if not overlap. Figure 1 represents the MDS visualization of the data for the subset of AHU parameters corresponding to the electricity consumption patterns/motifs: "*aaaa*" and "*bbbb*", and Fig. 2 represents the MDS visualization of the data of the same parameters corresponding to the energy consumption patterns: "*bcca*" and "*cccc*".

As shown in Figs. 1 and 2, the points are far away from each other. That means the AHU performance parameters (shown in Table 2) had significant changes over the years for all electricity consumption motifs ("*aaaa*", "*bbbb*", "*bcca*", and "*cccc*"). Since the AHU features analyzed in this study indicate system performance, detected change could be attributed to the degradation of the equipment, unless weather

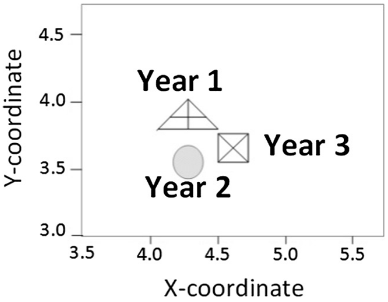


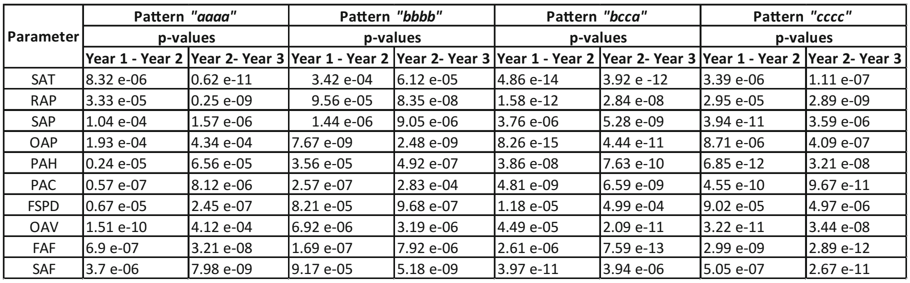
Fig. 3. MDS visualization for the weather data.

conditions change significantly. In order to ensure that the difference is not because of the weather conditions that change over the years, the authors checked if there is a significant change between the weather conditions in each year. It was concluded that the weather data does not differ significantly over the years, with the difference between points in both X and Y-coordinates being less than 0.5 (as shown in Fig. 3).

4 Statistical Analysis

The authors also conducted a statistical hypothesis testing in order to assure that the difference between the data of different years is statistically significant. Wilcoxon test was applied on each variable in the dataset (Table 2) over the years to detect if the change in the variables were significant. Wilcoxon test was mainly selected because of its applicability to the question investigated through hypothesis testing. The question investigated in this study was related to finding a difference between data measured over the years of Year 1, 2, and 3. The null hypothesis of Wilcoxon test was stated as "the median difference between pairs of observations (i.e., data of two consecutive years) is zero" [14]. Since Wilcoxon test is a two-tailed test, when the significance level is selected as 0.05, p values less than 0.025 results in rejection of the null hypothesis, implying that the difference between the data of two consecutive years is statistically significant.

Figure 4 illustrates p-values obtained from Wilcoxon Test across years for each of the selected feature. As shown in Fig. 4, all the p values obtained from the Wilcoxon Test is less than the significance level 0.025, implying that each of the AHU parameters in Fig. 4 experiences statistically significant behavioral changes between pairs of years (i.e., Year 1–2 and Year 2–3). This contradicts with the expectation that performance parameters should be configured similarly, when operating strategy of the AHU (electricity use patterns) is similar over the years. Therefore, the performance of the AHU analyzed in this study might have been affected by deterioration over time.



Parameter	Pattern *"aaaa"*		Pattern *"bbbb"*		Pattern *"bcca"*		Pattern *"cccc"*	
	p-values		p-values		p-values		p-values	
	Year 1 - Year 2	Year 2- Year 3	Year 1 - Year 2	Year 2- Year 3	Year 1 - Year 2	Year 2- Year 3	Year 1 - Year 2	Year 2- Year 3
SAT	8.32 e-06	0.62 e-11	3.42 e-04	6.12 e-05	4.86 e-14	3.92 e -12	3.39 e-06	1.11 e-07
RAP	3.33 e-05	0.25 e-09	9.56 e-05	8.35 e-08	1.58 e-12	2.84 e-08	2.95 e-05	2.89 e-09
SAP	1.04 e-04	1.57 e-06	1.44 e-06	9.05 e-06	3.76 e-06	5.28 e-09	3.94 e-11	3.59 e-06
OAP	1.93 e-04	4.34 e-04	7.67 e-09	2.48 e-09	8.26 e-15	4.44 e-11	8.71 e-06	4.09 e-07
PAH	0.24 e-05	6.56 e-05	3.56 e-05	4.92 e-07	3.86 e-08	7.63 e-10	6.85 e-12	3.21 e-08
PAC	0.57 e-07	8.12 e-06	2.57 e-07	2.83 e-04	4.81 e-09	6.59 e-09	4.55 e-10	9.67 e-11
FSPD	0.67 e-05	2.45 e-07	8.21 e-05	9.68 e-07	1.18 e-05	4.99 e-04	9.02 e-05	4.97 e-06
OAV	1.51 e-10	4.12 e-04	6.92 e-06	3.19 e-06	4.49 e-05	2.09 e-11	3.22 e-11	3.44 e-08
FAF	6.9 e-07	3.21 e-08	1.69 e-07	7.92 e-06	2.61 e-06	7.59 e-13	2.99 e-09	2.89 e-12
SAF	3.7 e-06	7.98 e-09	9.17 e-05	5.18 e-09	3.97 e-11	3.94 e-06	5.05 e-07	2.67 e-11

Fig. 4. Wilcoxon test results.

5 Discussions and Conclusion

In this paper, a methodology has been developed to quantify and visualize performance degradation of HVAC systems over the years using BASs data. The methodology consists of data pre-processing, motif discovery, feature selection, and visualization steps. This methodology has been applied to a case study of one of the AHUs of a smart building equipped with BAS. With this methodology, it is possible for facility operators to discover common system behaviors, such as the four common motifs that were identified as "*bcca*", "*aaaa*", "*bbbb*", and "*cccc*" for the studied AHU (where the scale represented increasing consumption from *a* to *c*).

With this methodology, it is also possible for facility operators to reduce the dimensionality of the AHU parameters and get a subset that directly affect the electricity consumption of that AHU. For example, ten AHU operational parameters were identified out of forty parameters as relevant to the AHU electricity use in the feature selection step. Resulting multi-dimensional dataset would then show how the set points defined for the parameters such as Supply Air Temperature, Fresh Air Flow, Return Air Pressure, etc. affect the consumption behavior of the AHU. For example, for the data set analyzed, the AHU performance parameters had significant changes over the years for all electricity consumption motifs *("aaaa", "bbbb", "bcca", and "cccc")* when weather data is controlled. The computational method provided in this paper provides an opportunity for facilities operators to get symbolic representation of the analyzed data corresponding to low, medium, and high energy consumption in buildings, and help operators to identify any changes in operations in major equipment corresponding to given energy profiles over the years. Such analysis is not possible with 24/7 streaming raw BAS data, which provides data on several HVAC equipment parameters with high frequency of sampling.

The methodology also helps effectively visualize the multi-dimensional dataset by transforming it into the Cartesian coordinate system in order to quantify and visualize the difference between datasets over the years (e.g., Years 1–3 in the case study). When facility operators look at large datasets with multiple dimensions, it is absolutely impossible to see the behavioral changes in the system over the years with well-known visualization plots. Multi dimension scaling used in this methodology helps operators to see the differences in system parameters effectively in two dimensions. Since MDS preserves inter-point distances between the data analyzed, if the points given by MDS are in close proximity with each other, then, the difference between dataset of different years are close to each other. For example for the studied data set, a quick glance at the MDS visuals shows that the data set differs over the years.

Future work includes evaluating the methodology over various types of HVAC system components, with ground truth data on system deteriorations. The current work is limited to a single case, where an extensive validation strategy is required to accurately attribute the behavioral changes over years to deterioration.

References

1. Katipamula, S., Brambley, M.R.: Methods for fault detection, diagnostics, and prognostics for building systems—a review, part I. HVAC&R Res. **11**(1), 3–25 (2005)
2. U.S. EIA: Annual Energy Review 2011. U.S. Energy Information Administration (2011). http://doi.org//EIA-1384(2011)
3. Zhang, R., Hong, T.: Modeling and simulation of operational faults of HVAC systems using energyplus. Proceedings of SimBuild, [S.l.], Aug 2016. http://ibpsa-usa.org/index.php/ibpusa/article/view/372. Accessed 27 Apr 2018
4. Brambley, M.R., Haves, P., McDonald, S.C., Torcellini, P., Hansen, D.G., Holmberg, D., Roth, K.: Advanced sensors and controls for building applications: market assessment and potential R&D pathways (No. PNNL-15149). Technical report, Pacific Northwest National Laboratory (PNNL), Richland (2005)
5. Wang, S., Wang, J.B.: Robust sensor fault diagnosis and validation in HVAC systems. Trans. Inst. Meas. Control **24**(3), 231–262 (2002)
6. Schein, J., Bushby, S.T., Milesi-Ferretti, N.S., House, J.: Results from field testing of air handling unit and variable air volume box fault detection tools. Technical report, NIST Interagency/Internal Report (NISTIR)-6994 (2003)
7. Beghi, A., Brignoli, R., Cecchinato, L., Menegazzo, G., Rampazzo, M., Simmini, F.: Data-driven fault detection and diagnosis for HVAC water chillers. Control Eng. Pract. **53**, 79–91 (2016)
8. Ahmad, M.W., Mourshed, M., Yuce, B., Rezgui, Y.: Computational intelligence techniques for HVAC systems: a review. Build. Simul. **9**(4), 359–398 (2016)
9. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Discov. **15**(2), 107–144 (2007)
10. Sutha, K., Jebamalar Tamilselvi, J.: A review of feature selection algorithms for data mining techniques. Int. J. Comput. Sci. Eng. **7**(6), 63 (2015)
11. Robinson, S.L., Bennett, R.J.: A typology of deviant workplace behaviors: a multidimensional scaling study. Acad. Manag. J. **38**(2), 555–572 (1995)
12. Gutierrez-Osuna, R., Nagle, H.T.: A method for evaluating data-preprocessing techniques for odour classification with an array of gas sensors. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **29**(5), 626–632 (1999)
13. Miller, C., Nagy, Z., Schlueter, A.: Automated daily pattern filtering of measured building performance data. Autom. Constr. **49**, 1–17 (2015)
14. Gehan, E.A.: A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika **52**(1–2), 203–224 (1965)

Vulnerability Distribution Model of Critical Infrastructures Based on Topological System Simulation

Xiaobo Yao[1], Chuanfeng Han[1], Qian Chen[1(✉)], and Lingpeng Meng[2]

[1] Department of Management Science and Engineering,
School of Economics and Management, Tongji University,
No. 1239 Siping Road, Shanghai 200092, People's Republic of China
chenqian19910610@hotmail.com

[2] China Institute of FTZ Supply Chain, Shanghai Maritime University, 1550
Haigang Ave, Pudong Xinqu, Shanghai 200120, People's Republic of China

Abstract. Urban critical infrastructures provide citizens with lifeline functions such as water, electricity and energy, etc. These interconnected infrastructure systems require reliable models for vulnerability measurement and topological controllability against usual disruptions and unusual hazards. This paper proposes a vulnerability distribution model to describe vulnerability distribution patterns in critical infrastructure system. To describe transmission of vulnerability between different infrastructure components or topological nodes, a vulnerability distribution network (VDN) is developed for simulation of negative impact on each node. The results are represented in a rasterized distribution diagrams by three metrics of vulnerability: the total number of effective topological nodes, the node's serviceability and the descent rate of coverage of infrastructural service. Then this model is applied to a case study of a local gas system and a local electric power system. Results show that a node's vulnerability and serviceability is closely related to the node's degree, especially the out-degree, while overall system's vulnerability is greatly affected by descent rate of coverage of each infrastructural service node. The model also generates probabilistic simulation graphs to show continuous vulnerability distribution in areas covered by the specified critical infrastructure systems. The graphic representation of VDN model results further helps infrastructure managers conduct route planning for transportation of hazard goods and optimize allocation of emergency resources.

Keywords: Critical infrastructure systems
Vulnerability distribution and assessment · Topological network simulation

1 Introduction

The critical infrastructure system, composed by a large number of interdependent physical facilities, technical units, operating processes and institutionalized control, provides necessary daily services for human society. Due to the increasing complexities of critical infrastructures, there would have been more sizeable amplification of effects

© Springer International Publishing AG, part of Springer Nature 2018

I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 498–515, 2018.
https://doi.org/10.1007/978-3-319-91635-4_26

of failure on the performance of these interconnected critical infrastructures [1]. The main challenge of management of critical infrastructures is what measures should be taken to enhance their robustness and functional reliability [2]. Questions have been highlighted when infrastructure managers and policy makers tried to seek the vulnerability distribution and changing patterns responsible for the propagating failure of interdependent critical infrastructures.

The state-of-the-art research achievements of dealing with analysis of the vulnerability of critical infrastructures are based on dynamical and time-varying networks, i.e. networks consisting of dynamical nodes with links that can change over time [3]. The treatment of vulnerability is often seen as stochastic measures of a set of random variables, including time to full system service resilience, cost of full system restoration [4]. Another network-based model, developed for European Commission, as a combined Systems Engineering and Dynamic Inoperability Input–output model (SE-DIIM), provides a complete framework for assessing the economic impact of critical infrastructure network failure on the national or regional level [5]. However, the existing methodologies of analyzing networked infrastructure systems mostly focus on network topological structural elements per se, which neglects the analysis of practical service-oriented network vulnerability in real application context. Seldom are there concrete studies on vulnerability distribution among interconnected and multiple different types of infrastructures. It is necessary to explain how vulnerability is distributed or propagated in a system of system and how serviceability is affected accordingly.

Even if vulnerability factors are normally very-infrastructure specific and hazard dependent, generic factors can be defined to account for a variety of probabilistic calibration of vulnerability, such as buffer capacity, adaptability and redundancy [6]. Vulnerability can also be represented by maturity level of critical infrastructure protection preparedness, through the root-cause in a Delphi survey [7]. However, these measuring indicators are subjective, making it hard to quantify and simulate them in a highly networked system accurately. Considering the networked attributes of critical infrastructures, several topological metrics should be included in the evaluation of the vulnerability. The evaluation results in turn provide implications for managers to plan transportation and optimize emergency resource allocation for vulnerable areas.

In order to give an overall assessment of vulnerability of a single-type infrastructure and multi-service infrastructure form a serviceability perspective, this paper proposes a vulnerability distribution model based on topological system simulation. The proposed model will concentrate on service levels of critical infrastructures in terms of the vulnerability distribution impact on each topological node. This will allow infrastructure managers and policy makers to be aware of where the potential risks transmit within the networked infrastructure system, thereby to take actions to reduce negative impact or failures caused by risks.

2 Vulnerability Distribution Network (VDN)

Vulnerability, as brittleness, is the diminished capacity of an individual or group to anticipate, cope with, resist and recover from the impact of a natural or man-made hazard (defined by IFRC [8]). When extending this definition to networked systems of the critical infrastructures, it is obvious that each infrastructure component is prone to cause the cascade failure impact on the service provided by the overall system. Multiple infrastructure components compose a topological network that represents an infrastructure system, and each infrastructure component is viewed as a node in this network system. For example, in an infrastructure system of electric power service, random failure of a small transformer substation will lead to the cascade failure of other stations, even to the breakdown of power service. The areas covered by the infrastructure system and the multiple infrastructure components can be rasterized like an image. The idea of rasterizing the area helps to analyze the connectivity or vulnerability between different nodes and points. In this paper, the vulnerability of infrastructure system is defined as the diminished capacity of critical infrastructures to provide continuous services after negative interferences.

Based on topological characteristics of critical infrastructure system and their coupling relationships, vulnerability is able to distribute among different nodes. Then these nodes and the vulnerability distribution route in between comprise a vulnerability distribution network (VDN). The vulnerability distribution network is, generally speaking, a topological model to describe how all of the included infrastructure services are transmitted and influenced. Assuming F_i is the service capacity (i.e., serviceability) of a service node i, and VDN_{ij} is an edge from node i to node j, then the vulnerability distribution network N_{VDN} is a collection of nodes and edges, shown in (1). In a particular VDN, the edge between two nodes means a direct vulnerability transmission relationship, shown as a directed edge in a VDN graph. Vulnerability distribution pattern is shown in Fig. 1. This distributing process, or transmitting process, is either unidirectional or bi-directional.

$$N_{VDN} = \{F_i, VDN_{ij} | i,j = 1, \ldots, N, i \neq j\} \tag{1}$$

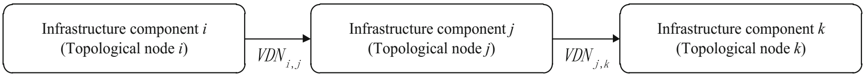


Fig. 1. Vulnerability distribution pattern.

The vulnerability of an individual infrastructure component is described with a variable set $(\lambda, \Delta t)$. λ represents the threshold value when the topological node can no longer provide continuous service, meaning that the node's serviceability is collapsed after its capacity is diminished to λ. Δt is the delay time contained before the vulnerability starts to distribute. The vulnerability distribution, as an edge in a VDN, is described as $VDN_{ij} = (P_{ij}, t_{ij}, s_{ij})$. P_{ij} represents the probability that the vulnerability of node i is distributed to node j. To simplify the model, the probability of vulnerability distribution from node i to node j, is assumed to be an exponential function of the physical distances between these nodes. t_{ij} is the distributing time from node i to node j, and s_{ij} means the final state of node j. If $s_{ij} = 1$, the service of node j is interrupted. If $s_{ij} = 0$, node j performs continuous service. The value of the set {0, 1} is viewed as an activated state of a node. To simplify the model calculation, only edges are left when s_{ij} equals 1. The detailed properties of a particular VDN_{ij} are given in (2).

$$0 \leq P_{ij} \leq 1, P_{ik} = P_{ij} * P_{jk}, t_{ij} = t_{ij} + t_j + t_{jk} \tag{2}$$

Considering the complex performance of interconnected infrastructure systems, the node degree, which is the number of connections (edges in the *VDN* graph) each node has to other nodes, usually possesses a value larger than one. For a single node, the number of edge arrowhead ends to it is called the in-degree, and the number of edge tail ends to it is called out-degree. When the in-degree of a node i is K_i, the vulnerability of this node to have serviceability inhibited is determined by $2(K_i - 1)$ activated states, since as aforementioned each node has two activated state $\{0,\ 1\}$. In fact, the vulnerability distribution network can simulate the activated state of all the related nodes in order to reach a final vulnerability status of the infrastructure systems.

3 Vulnerability Distribution Model (VDM)

In order to address the problem of homogeneous topological analysis for infrastructure functionality, the different states of vulnerability distribution of a single-service infrastructure system network (e.g. a single gas service infrastructure composed by subordinate infrastructure elements) and a multi-service infrastructure (e.g. a system composed by gas infrastructure and electric power infrastructure) are considered. This paper proposes two separate vulnerability distribution models (VDM) for analysis; VDM for single-service infrastructure system and VDM for multi-service infrastructure system. Both of the VDM conduct simulation using the probability properties of VDN distribution in Sect. 2. VDM for multi-service infrastructure system is a multi-dimensional combination of two or more single-service infrastructure system, while a single node's in-degree and out-degree expanded considerably through an extension of vulnerability distribution edges.

3.1 VDM for Single-Service Infrastructure System

It contains three metrics for vulnerability simulation of a single-service infrastructure system: the total number of effective topological nodes $V_{j|N}$, the nodes' serviceability $V_{j|S}$ and the descent rate of coverage of infrastructural service $V_{j|C}$. If we consider node j as the source node of vulnerability distribution that is able to have negative impact on its interconnected node i, t_0 as the initial time of vulnerability distribution, T_{ij} as the distributing time, g_i as the vulnerability status of node i, f_i as the serviceability of node i, then $V_{j|N}$ and $V_{j|S}$ are simply described in (3) and (4). In order to ensure the stability and extensity of the infrastructure service, overlapping of node's service area usually exist in the overall network. Therefore, $V_{j|C}$ provides more insightful information about the system's serviceability and vulnerability rather than an individual node's performance. As aforementioned, rasterization of the network area to have multiple points is a rational way to measure $V_{j|C}$ accurately. For example, a rectangular area of 12 km^2

can be divided in 300 * 400 rasterized sub areas, with 300 * 400 rasterization points. Hence, the calculation formula for $V_{j|C}$ is given in (5), where C_k denotes the serviceability of a rasterization point k. Similar to a node's activated state, C_k only has two values: 1 (service available from neighboring nodes) and 0 (no service provided by neighboring nodes at all).

$$V_{j|N} = \frac{\sum_i (g_i(t_0) - g_i(t_0 + T_{ij}))}{\sum_i g_i(t_0)} \tag{3}$$

$$V_{j|S} = \frac{\sum_i \left(g_i(t_0) - g_i\left(t_0 + T_{ij}\right)\right) * f_i}{\sum_i g_i(t_0) * f_i} \tag{4}$$

$$V_{j|C} = \frac{\sum_k (C_k(t_0) - C_k(t_0 + T_{ij}))}{\sum_k C_k(t_0)} \tag{5}$$

3.2 VDM for Multi-service Infrastructure System

Because a complex network has a "system of system" in nature, it is necessary to clarify the vulnerability distribution of critical infrastructures composed by various categories of single infrastructure system. According to President's Commission on Critical Infrastructure Protection Report [9], the basic lifeline to support a national society includes critical infrastructures: transportation, oil and gas production and storage, water supply, emergency services, government services, banking and finance, electrical power and telecommunications. It is obvious that these infrastructures also have intricacy and interdependent topological relations [10]. For example, electrical power system satisfies a huge demand of electricity from almost all the other critical infrastructure systems.

Featuring a "system of system" structure, VDM is a multi-dimensional topological model. In fact, it is a multi-dimensional combination of two or more single-service infrastructure system, while a single node's in-degree and out-degree expanded considerably through an extension of vulnerability distribution edges. Adding one type of infrastructure leads to an addition of model dimension. Considering a multi-dimensional representation of VDM, three metrics: the total number of effective topological nodes, the nodes' serviceability and the descent rate of coverage of infrastructural service become $\overrightarrow{V_j}(t_0) = (V_j^1, \ldots, V_j^k, \ldots, V_j^D)$ in D-dimension space. Assuming the node and rasterization point of multi-dimensional model is unified, therefore the calculation and simulation of the three metrics follows the same formulas as the single-service VDM, while showing multi-dimensional attributes. In reality, the demand level and functional significance of different types of infrastructure would vary, it is necessary to define the weight of each element of $\overrightarrow{V_j}(t_0)$. The weight $\vec{W} = (W_1, \ldots, W_D)$ should be assigned according to a specific service time frame, the location of the critical infrastructure, and its service impact on the local society. By combining the vulnerability distribution metrics and appropriate weights, an integrated vulnerability state of the big critical infrastructure system V is derived:

$$V = \vec{W} * \vec{V} \tag{6}$$

3.3 Calculation of Vulnerability Values Based on VDM

VDM gives a representation of vulnerability distributing patterns considering the value on the node level, for example, the vulnerability of a power plant itself rather than its serviceable area. In real situations, vulnerability of an infrastructure is determined by the level of spectrum of services (i.e., the serviceable area with a certain radius range) it can provide. The service spectrum usually conforms design specifications of that infrastructure. Taking the electrical power system as an example, a 220 kV power plant should be able to provide electricity within 100–300 km radius, while a 10 kV terminal station should be able to provide electricity within kilometers radius. In VDM, the serviceable areas are all simplified as circular areas with a specific radius that is determined by design specifications of a specific type of infrastructure. A VDM example showing the level of spectrum of service is given in Fig. 2.

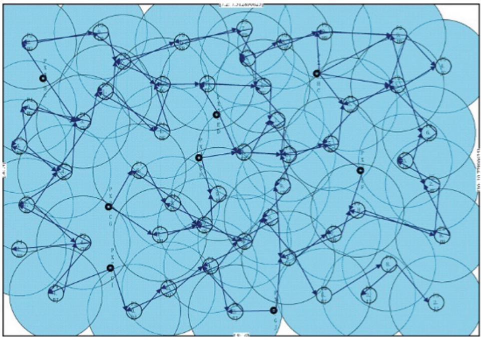


Fig. 2. A VDM example showing the level of spectrum of service.

With respect to the descent rate of coverage of infrastructural service, it is critical to identify the number of its constituting nodes and the level of spectrum of services each node provides. As aforementioned, a network area can be divided into a number of rasterized sub areas. Thus, the calculation of vulnerability value of a type of infrastructure or of an interconnected rasterization subs areas will be determined by the number of infrastructure nodes and the level of spectrum of its service that cover that cluster of grids. If a rasterization point is labeled as $\overrightarrow{r_i}$ that is service-covered by infrastructure i, the probability of vulnerability distribution from node i to the rasterization point is generated by a function $p(\overrightarrow{r_i})$. That is, the probability of vulnerability distribution is only associated with relative distance between two points.

The vulnerability on the rasterization point refers to the negative impact on normal functionality of an infrastructure, and its value is determined by $p(\overrightarrow{r_i})$. It is also noted that the rasterization point is affected by more than one infrastructure component when two or more infrastructure nodes radiate overlapping serviceable areas. Therefore, for the same rasterization point, it will also receive probability of vulnerability distribution from node *j*, meaning that its vulnerability will be affected by $p(\overrightarrow{r_i})$. In order to mathematically compute the final value of vulnerability at any rasterization point or topological node, (7) is used when considering the overlapping coverage of service from both infrastructure node *i* and *j*. When the number of rasterization points reaches infinity great, the vulnerability value becomes a continuous function.

$$V_{ij}(\overrightarrow{r_i}) = p_i(\overrightarrow{r_i}) * (1 - p_j(\overrightarrow{r_j})) * V_i + (1 - p_i(\overrightarrow{r_i})) * (\overrightarrow{r_j}) * V_i + p_i(\overrightarrow{r_i}) * p_j(\overrightarrow{r_j}) * V_{ij} \quad (7)$$

4 Vulnerability Distribution Network Simulation Based on VDM

4.1 Simulation Platform and Processes

Most of the research achievements of vulnerability modeling and simulation are within the dynamic simulation categories. A well-known one is CIP/DSS (Critical Infrastructure Protection) system [11] developed by Los Alamos National Lab in North America. Dynamic processes, such as delivery of fuel for repair vehicles, are represented in the CIP/DSS infrastructure sector simulations by differential equations, discrete events, and codified rules of operation. Fort Future [12], developed by United States Army Corps of Engineers (USACE), is a vulnerability simulation system emphasizing on the collaborative behavior of different infrastructure participants. These simulation systems provide basic concepts for simulation processes in this research, while the distinctive feature of our simulation platform is its ability to reveal the vulnerability distribution patterns both on a node level and on a serviceable area level.

Microsoft Visio is applied to simulate vulnerability based on customized data structure (e.g., the rasterization point list “cCell”, single node list “cNode”, etc.) defined on this simulation platform. These initial data of topological nodes of predefined infrastructure systems are stored in Microsoft Excel files. VDM is computed in Visio when it receives and reads the initial data, then it uses the visually display function to show the graphic simulation results. For each initial scenario, vulnerability distribution simulations are carried out repeatedly using Monte Carlo technics. Raw data “cNode” from an example VDN is given in Table 1 and its corresponding simulation result is given in Table 2. The graphic simulation results are differentiated by thickness of colors in topological area, which are also called the “vulnerability cloud”. The thicker distribution cloud implies more vulnerability. The vulnerability distribution network simulation process consist of eight elements, which is shown in Fig. 3.

The simulation platform includes three main modules in order to materialize the simulation process. The first module is “front end display module”, which uses VISIO user interface to display the shape of infrastructure network and its rasterization sub areas

Table 1. Raw data “cNode” from an example VDN

Node i name	Node j name	T_{ij}	$p_i\left(\overrightarrow{r_i}\right)$	s_{ij}
GX4	G6	0.0002	1	0.5
G6	G11	0.0002	1	0
G6	G26	0.0001	1	0.7
GX4	G13	0.0001	1	0.5
G13	G26	0.0001	1	0.7
G13	G22	0.0001	1	0
…	…	…	…	…
G5	G7	0.0001	1	0
GX1	G10	0.0001	1	0.5
GX1	G27	0.0002	1	0.5
GX1	G21	0.0002	1	0.5
G21	G23	0.0001	1	0
G10	G21	0.0003	1	1

Table 2. Simulation results of the example VDN

In-degree	Number of topological nodes	V_{ij}			
		Min value	Max value	Average value	Variance
0	8	0.00083 0.00054 0.00001	0.00583 0.00701 0.00664	0.00250 0.00303 0.00249	0.001816 0.002236 0.002343
1	5	0.00083 0.00095 0.00020	0.00333 0.00381 0.00365	0.00167 0.00190 0.00151	0.001054 0.001204 0.001333
2	25	0.00083 0.00095 0	0.00250 0.00333 0.00345	0.00051 0.00146 0.00068	0.000512 0.000697 0.000810
3	16	0.00083 0.00095 0	0.00417 0.00476 0.00414	0.00125 0.00172 0.00068	0.000977 0.001179 0.001184
4	6	0.00083 0.00095 0	0.00250 0.00314 0.00154	0.00125 0.00180 0.00050	0.000636 0.000878 0.000549

as well as the service status of each infrastructure node. The second module is “back end VBA I/O module”, responsible for data interaction between VISIO and Excel. The third module is “back end VBA simulation module”, which carries out all the computations with respect to vulnerability calculations. An example function to determine the end state of each node is given in Appendix. When a simulation is initialized, all the parameter settings (already written in Excel files) will be read through VBA I/O module and transferred to the VBA simulation module. After the simulation is completed, the output data will in turn go back to Excel files through the same VBA I/O module.

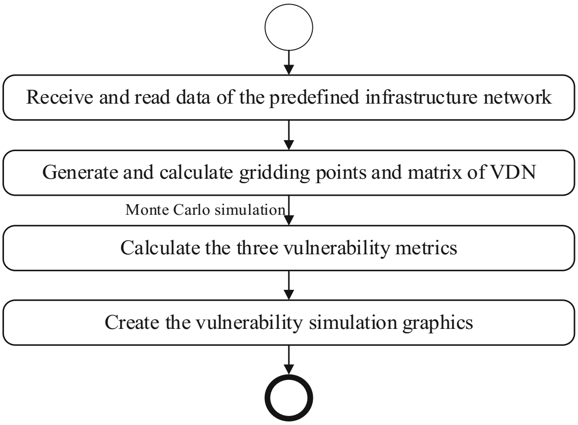

Receive and read data of the predefined infrastructure network
Generate and calculate gridding points and matrix of VDN
Monte Carlo simulation
Calculate the three vulnerability metrics
Create the vulnerability simulation graphics

Fig. 3. The vulnerability distribution network simulation processes.

The simulation results (i.e., output data) are displayed as graphic representations of vulnerability clouds with the help of Origin software, and shown on the VISIO front end user interface.

4.2 Simulation Method and Required Data Structure

Similar to those well-known vulnerability simulation platforms, this research applies the discrete event simulation method and event chain based method to conduct the simulation. Discrete event simulation (DES) is the process of codifying the behavior of a complex system as an ordered sequence of well-defined events. Each event occurs at a particular instant in time and marks a change of state in the system. Events are scheduled in most discrete systems by event calendar or event chain. The events generated by the model join a time-ordered list. As the simulation clock is advanced using the periodic scan approach, the event calendar is scanned for events that should have occurred in the last time unit [13]. In VDM, the event chain refers to a series of five meta-event, shown in Fig. 4.

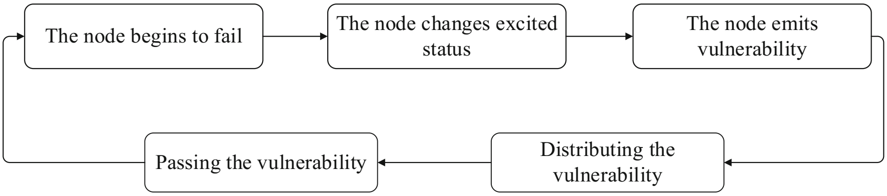

The node begins to fail
The node changes excited status
The node emits vulnerability
Distributing the vulnerability
Passing the vulnerability

Fig. 4. The event chain of VDM.

In order to perform the discrete event simulation for cascading effect of failure in infrastructure network, the rules of simulation are predetermined. That is, the status of each infrastructure node or rasterization point will retain unless it reaches an activated state, which activate the vulnerability distribution to the interconnected nodes or rasterization points. The cascading effect should be validated within this network.

According to the vulnerability distribution model and discrete event simulation method, the primary data structure is consisted of cCell data, cNode data, and cVD data. The cCell data includes all the information of rasterization points that are used for dividing overall-covered infrastructure network area, such as coordinate x, coordinate y, area, design specification related to serviceability of node it belongs to, the degree of grid coverage, and the vulnerability of the point. The cNode data includes all the information of topological nodes that represent the infrastructure properties, such as location, service type, service capability, level of spectrum of service, threshold value of vulnerability distribution to change activated state, the number of rasterization points each node covers, the covered rasterization point list and the vulnerability of each node. With respect to simulation processes, cNode also contains time information such as time between discrete event steps and number of steps. The cVD data includes configuration information about the distribution model, such as the connection type between nodes, whether connected nodes are crossing different infrastructure systems, connected nodes No., and probability of vulnerability distribution on each topological edge.

Besides the primary data structure, there is data describing the queuing behavior of discrete events. A set of data tuples are used to store the information for event queuing, such as the numbering of event, delay of event for each node, the top event within a queue, whether to update the queuing status, etc. Only by coupling with all these data, the simulation can be conducted to generate the graphic vulnerability pattern (i.e., distribution cloud).

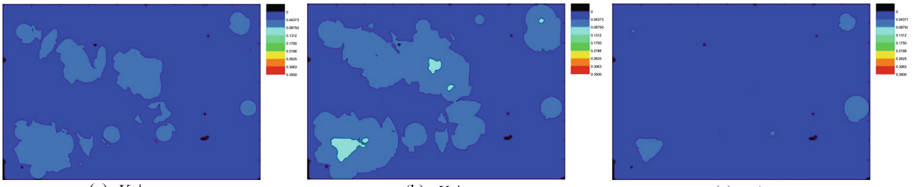
5 Implementation of Case Study

5.1 VDM for Single-Service Infrastructure System (Gas System)

In this example, a gas infrastructure system consists of 30 service nodes and 35 paths of pipeline is analyzed, in which each node is providing service for a circular area with a predefined radius. In the simulation program, the area is divided into 100 * 150 grids, and the gas infrastructure system provides a 99.79% coverage initially. The simulation results are shown in Fig. 5. In terms of the node's serviceability, results show bulky vulnerability in left side area with more in-degree vulnerability. While concerning the total number of effective topological nodes and the descent rate of coverage of infrastructural service, the vulnerability distribution mainly concentrates on key nodes of gas infrastructure system in a small area.

5.2 VDM for Single-Service Infrastructure System (Electrical Power System)

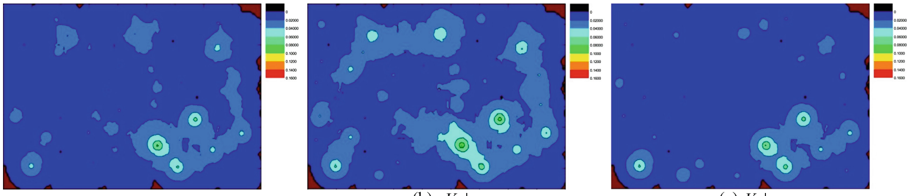
In this example, an electrical power infrastructure system consist of 60 nodes and 76 paths of electric power line is analyzed. The topological nodes represent 8 power plants and 52 transformer stations. The system covers the same area with the gas system.



(a) $V_j|_N$ (b) $V_j|_S$ (c) $V_j|_C$

Fig. 5. Graphic simulation results of vulnerability distribution of gas system.

After the area is divided into 100 * 150 grids in the simulation program, and the electric system provides an electric coverage of 98.35%. The simulation results are shown in Fig. 6, and according to (a) and (b), the vulnerability distributions, in terms of the total number of effective topological nodes and the node's serviceability, show similarly simple variation tendency with a focus on key nodes. However, the descent rate of coverage of infrastructural service performs a complex shape of vulnerability distribution, in a larger area of vulnerability. It proves that, in a single electric power infrastructure system, the node activated state is influenced by number of rasterization points and nodes in a specific area.



(a) $V_j|_N$ (b) $V_j|_S$ (c) $V_j|_C$

Fig. 6. Graphic simulation results of vulnerability distribution of electric power system.

5.3 VDM for Multi-service Infrastructure System (Correlated Gas-Electrical Power System)

In a two-dimensional multi-infrastructure gas-electric power system, the weight of gas service is set as 0.4 and of the electric power service as 0.6, $\vec{W} = (0.6, 0.4)$, based on expert scores. The topological representation of the network structure (see Fig. 7, where the red edge indicates gas infrastructure out-degree, the blue edge indicates electric power infrastructure out-degree, and the black lines simply represent traffic route in the area) is formulated by linking the gas vulnerability distribution node with its interdependent electric power node. For example, a gas station GGX-1 has power supply from a transformer substation EP-9, meanwhile the gas station GX-1 provides gas service to support power station PX-2. The graphic simulation results are shown in Fig. 8.

In terms of the three vulnerability metrics, the graphic representations are similar in the shapes as the corresponding ones in single-infrastructure simulation. Compared to single-infrastructure system, it is obvious that multi-infrastructure demonstrates significantly bigger vulnerability and wider vulnerability distribution area. However, the

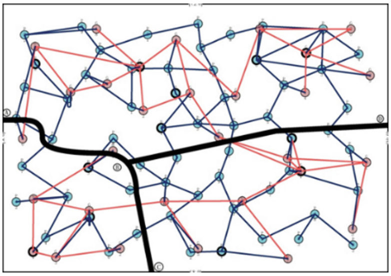
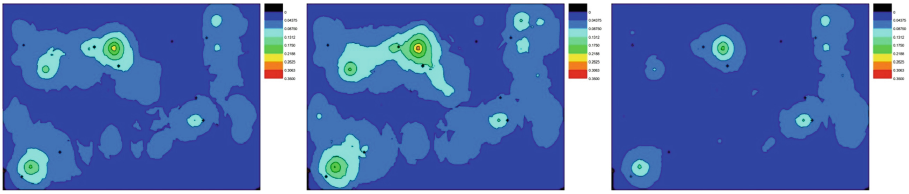


Fig. 7. Topologic representation of correlated gas-electricity system. (Color figure online)



(a) $V_j|_N$ (b) $V_j|_S$ (c) $V_j|_C$

Fig. 8. Graphic simulation results of vulnerability distribution of correlated gas-electricity system.

single-infrastructure simulation results of electric power do not manifest a significant difference as gas when compared to multi-infrastructure vulnerability performance. Gas stations do not have a considerable impact on electric power infrastructure, where the vulnerability mostly comes from electrical elements themselves. Therefore, it reveals that the vulnerability distribution in a correlated multi-infrastructure system will have non-linear relationships with its single dimensional vulnerability of single-infrastructure.

Nevertheless, in this case, due to additional vulnerability distribution in-degree from electric power system to gas system, the gas system as a vulnerability receiver will significantly increase its operational risk. An implication for infrastructure managers is that it is critical to take into account the change of vulnerability distribution caused by components' interdependencies in a multi-infrastructure system and then describe the vulnerability on each service dimension.

5.4 VDM During Different Time Periods

In real practice, the serviceability of an infrastructure node is affected by the time periods during a day (i.e., 24 h). It is necessary to differentiate the demand level and functional significance of each node during day time and night time. This analysis based on time is especially critical when infrastructure managers are involved in traffic route planning for transportation of hazard goods. Therefore, the weights can be re-assigned to conduct vulnerability simulations using the same expert scoring method. 24 h can be divided into four main time periods; morning time (04:00–08:00), daytime

(08:00–18:00), dusk time (18:00–22:00), and night time (22:00–04:00). The weights assigned for different time periods are shown in Table 3. After running the simulation with new weights, the graphic simulation results of vulnerability distribution clouds are presented in Fig. 9.

Table 3. The weights assigned for different time periods.

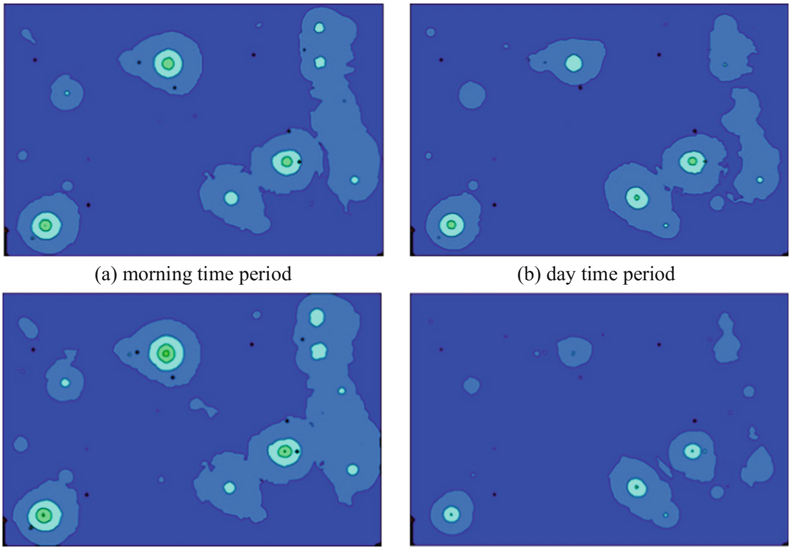
Time periods	Weights	
	Gas system	Electrical power system
Morning time	0.5	0.5
Daytime	0.3	0.7
Dusk time	0.6	0.4
Nighttime	0.2	0.8

The basic requirement of route planning for transportation of hazard goods is to avoid or reduce the transportation in highly vulnerable areas during the highest vulnerable time periods. The most vulnerable time periods for correlated gas-electricity system are morning time and dusk time. This is in line with the fact that transportation and infrastructure services are busy within this time frame. Therefore, infrastructure managers are suggested to ban the transportation of hazard goods during morning and dusk time. It is less risky to transport hazard goods during daytime and nighttime. Nevertheless, careful considerations should be given to route planning in order to differentiate daytime and nighttime. Due to the vulnerability distribution clouds, the route planning considering vulnerability of correlated gas-electricity system is shown in Fig. 10. During daytime, the right part of the region (dashed red line denoting the banned route in (a)) should ban the transportation because of high vulnerability of that rasterization area. But it is possible to open the same route during nighttime. In case of emergency events caused by leaking of hazard goods, it is suggested that emergency resources should be allocated on the right bottom part of the region due to the non-negligible vulnerability (as shown in (b)). In this way, vulnerability distribution clouds also help emergency departments to plan and allocate emergency resources.

6 Discussion

Unlike the traditional topological analysis focusing on homogeneous nodes, this research is characterizing the different serviceability of infrastructure components (i.e., topological nodes). To sum up, the highlights of this research are listed as follows:

1. The concept of vulnerability and vulnerability distribution (i.e., transmission and propagation) are reintroduced, and the critical infrastructure systems at certain levels are modeled as a vulnerability distribution network model which emphasizes on the distribution of services.



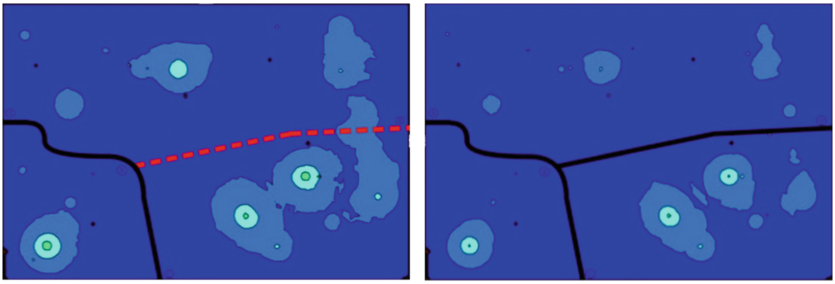
(a) morning time period

(b) day time period

(c) dusk time period

(d) night time period

Fig. 9. Graphic simulation results of vulnerability distribution of correlated gas-electricity system during different time periods.



(a) daytime route planning

(b) nighttime route planning

Fig. 10. Route planning considering the vulnerability distribution during daytime and nighttime. (Color figure online)

2. Service capability and service vulnerability are defined to measure the potential impact on system function and service coverage by topological simulation when experiencing collapse event of each node. Based on service coverage the vulnerability is expanded to rasterized distribution to improve computing efficiency.
3. For correlated infrastructure systems, unlike traditional network-based vulnerability analysis, VDM for all service dimensions are combined into an integrated cloudlike vulnerability distribution pattern.

4. The vulnerability distribution model and graphic representation as vulnerability clouds help guide infrastructure managers and emergency departments conduct route planning and optimize the allocation of emergency resources.

However, this research does not distinguish the level of rasterization of area. More rasterization points should give more reliable simulation results of vulnerability distribution. The demanding level of each rasterization point is not considered, while only the design specification of service capacity of each infrastructure node is analyzed on the covered rasterization areas. Geographic information system should be furthered incorporated into the simulation in order to add precision to route planning and emergency resource allocation.

7 Conclusion

An increasing number of risk management and disaster preparedness plans have been developed in recent years, however, only limited attention was given to the risk and vulnerability distribution patterns for critical infrastructures with a systematic perspective. This paper proposes a vulnerability distribution network and its analytical topology model for single-infrastructure and multi-infrastructure systems. Three metrics are defined to represent the vulnerability distribution characteristics. Through simulation, the interdependent relationship of components of the critical infrastructure system is evaluated effectively and show clearly in vulnerability simulation graphs.

The single-service infrastructure simulation results show that the descent rate of service coverage has more impact in a wider vulnerability distribution area in electric power infrastructure system. In the multi-service infrastructure system, it is significant that multiple single infrastructure systems will be interconnected to generate more out-degree and hence, increasingly complex vulnerability distributions. The vulnerability distribution model should be applied both to single and multiple critical infrastructure systems so that decision makers will know the forming process and distribution routes of venerability. Therefore, appropriate mitigation strategies and resource allocation plans are made to address corresponding problems. However, the model proposed is abstract and only on a theoretical basis. Future efforts should be given to improve vulnerability distribution estimated to be integrated with physical object locations.

Appendix: An Example Function to Determine the End State of Each Node

```
 'get inbound link info: end status
    Public Function sGetLinkIn(ByVal iLinkFromX As Integer) As Single
      sGetLinkIn = 0
      For I = 1 To iLinkFroms(0)
        If iLinkFroms(I) = iLinkFromX Then Exit For
      Next I
      If I <= iInDegree Then
        sGetLinkIn = REndStates(I)
      End If
    End Function

        Public  Function  sGetLinkInByNodeIndex(ByVal  iLinkFromIndex
As Integer) As Single
          If I <= iInDegree Then
            sGetLinkInByNodeIndex = REndStates(iLinkFromIndex)
          Else
            sGetLinkInByNodeIndex = -1
          End If
        End Function

        'sGetLinkInResult By the nodeindex of source Node
        Public  Function  sGetLinkInByLinkNode(ByVal  iLinkFromNode  As
Integer) As Single
          For I = 1 To iInDegree
            If iLinkFroms(I) = iLinkFromNode Then Exit For
          Next I
          If I <= iInDegree Then
            sGetLinkInByLinkNode = REndStates(I)
          Else
            sGetLinkInByLinkNode = -1
          End If
        End Function

    Public  Function  sAddLinkIn(ByVal  iLinkFromX  As  Integer,  rEndStateX
As Double) As Integer
      Dim I As Integer, J As Integer
      'check the existed links
      If iLinkFroms(0) = 0 Then
        I = 1
        iLinkFroms(I) = iLinkFromX
        REndStates(I) = rEndStateX
        sAddLinkIn = I
        iLinkFroms(0) = I
        iInDegree = iLinkFroms(0)
        rSingleLinkEndStatus = rEndStateX
        Exit Function
      End If
```

```
        For I = 1 To iInDegree
           If iLinkFromX = iLinkFroms(I) Then
              'if exist, then update it or delete it
              iLinkFroms(I) = iLinkFromX
              REndStates(I) = rEndStateX
              sAddLinkIn = I
              Exit For
           End If
        Next I
        'search for the right position to insert
        For I = 1 To iLinkFroms(0)
           If iLinkFromX > iLinkFroms(I) Then Exit For
        Next I
        'insert : expand the array'
        iInDegree = iInDegree + 1: iLinkFroms(0) = iInDegree
        ReDim Preserve iLinkFroms(0 To iInDegree)
        ReDim Preserve REndStates(0 To iInDegree)
        For J = iInDegree To (I + 1) Step -1
           iLinkFroms(J) = iLinkFroms(J - 1)
           REndStates(J) = REndStates(J - 1)
        Next J
        iLinkFroms(I) = iLinkFromX
        REndStates(I) = rEndStateX

        iInDegree = iLinkFroms(0)
        sAddLinkIn = I

     End Function
```

References

1. Chiuso, A., Fortuna, L., Frasca, M., Rizzo, A., Schenato, L., Zampieri, S. (eds.): Modelling, Estimation and Control of Networked Complex Systems. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03199-1
2. Xu, T., Masys, A.J.: Critical infrastructure vulnerabilities: embracing a network mindset. In: Masys, A.J. (ed.) Exploring the Security Landscape: Non-traditional Security Challenges. ASTSA, pp. 177–193. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27914-5_9
3. Setola, R., De Porcellinis, S., Sforna, M.: Critical infrastructure dependency assessment using the input–output inoperability model. Int. J. Crit. Infrastruct. Prot. **2**(4), 170–178 (2009)
4. Baroud, H., Ramirez-Marquez, J.E., Barker, K., Rocco, C.M.: Stochastic measures of network resilience: applications to waterway commodity flows. Risk Anal. **34**(7), 1317–1335 (2014)
5. Jonkeren, O., Azzini, I., Galbusera, L., Ntalampiras, S., Giannopoulos, G.: Analysis of critical infrastructure network failure in the European Union: a combined systems engineering and economic model. Netw. Spat. Econ. **15**(2), 253–270 (2015)
6. Eidsvig, U., Uzielli, M., Vangelsten, B.V.: Approaches for assessment of vulnerability of critical infrastructures to weather-related hazards. In: EGU General Assembly Conference Abstracts, vol. 18, p. 7722 (2016)

7. Karabacak, B., Yildirim, S.O., Baykal, N.: A vulnerability-driven cyber security maturity model for measuring national critical infrastructure protection preparedness. Int. J. Crit. Infrastruct. Prot. **162**, 1494–1504 (2016)
8. Burton, C.: IFRC Vulnerability and Capacity Assessment Training Review: 13 November–20 December. International Federation of the Red Cross Red Crescent Societies, Geneva (2007)
9. Ellis, J., et al.: Report to the President's Commission on Critical Infrastructure Protection. No. CMU/SEI-97-SR-00333. Software Engineering Institute, Carnegie-Mellon University, Pittsburgh (1997)
10. Lindbom, H., Tehler, H., Eriksson, K., Aven, T.: The capability concept–on how to define and describe capability in relation to risk, vulnerability and resilience. Reliab. Eng. Syst. Saf. **135**, 45–54 (2015)
11. Conrad, S.H., LeClaire, R.J., O'Reilly, G.P., Uzunalioglu, H.: Critical national infrastructure reliability modeling and analysis. Bell Labs Tech. J. **11**(3), 57–71 (2006)
12. Case, M.P., Smith, W., Grobler, F.: Fort future: modeling and simulation for collaborative multi-criteria decision support. In: Computing in Civil Engineering, pp. 1–10 (2005)
13. Pooch, U.W., Wall, J.A.: Discrete Event Simulation: A Practical Approach, vol. 4. CRC Press, Boca Raton (1992)

Component-Based Machine Learning for Energy Performance Prediction by MultiLOD Models in the Early Phases of Building Design

Philipp Geyer(✉), Manav Mahan Singh,
and Sundaravelpandian Singaravel

KU Leuven, Kasteelpark Arenberg 1 - box 2431, 3001 Leuven, Belgium
p.geyer@kuleuven.be

Abstract. The application of building information modeling (BIM) in early design phases requires the support of different levels of detail (LOD). This allows scaling to be supported as an important activity of designing. Furthermore, to achieve well-performing solutions in terms of energy efficiency, it is necessary to consider energy performance in early design stages. Therefore, this paper presents a multiLOD modeling approach for the early phases of building design that integrates energy performance prediction based on component-based machine learning (ML) using artificial neural networks (ANN). A model structure with three adaptive LOD definitions is proposed to support the design process by a digital model that supports flexible scaling back and forth. By linking the ML models to the elements in this structure, components are formed that support quick and flexible modeling and energy performance prediction in the early building design process. The transformation rules flexibly link the ML components to all LOD. This approach was illustrated and validated by a test case with a medium-sized office building. The early design states of the case were reconstructed for the application of the method. For validation purposes, the results of the ML predictions for 60 different design configurations were compared to those of a conventional parametric full-detail simulation model. This comparison showed that the average error was no higher than 3.8% for heating and 3.5% for cooling.

Keywords: Multi-level-of-detail modeling
Component-based machine learning · Early design phase
Building information modeling (BIM)

1 Introduction

The building industry is increasingly using the benefits of building information modeling (BIM) [1, 2] to generate and manage building information during evolving and iterative design processes [3]. However, the available representation and product modeling, such as the Industry Foundation Classes (IFC) [4], mostly focus on the later planning and construction phases. Early design phases, in which key decisions about building performance (such as energy efficiency) are made, are not well addressed by BIM representation.

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 516–534, 2018.
https://doi.org/10.1007/978-3-319-91635-4_27

Therefore, the development of BIM representations for early phases of design is highly relevant for design and should be addressed by researchers, which motivated the present research [5]. At the core of this study are multi-level-of-detail (multiLOD) models that enable designers to switch between the LOD of the model that are kept consistent, allowing the development of a design at different scales.

Performance plays an important role in BIM, especially in early design phases, in terms of the present demand for energy efficiency and sustainability. Major decisions about the performance of a building design are made in early design phases. BIM, especially the multiLOD approach, has the potential to support effective informed design-decisions. However, the current energy performance assessment methods are not available in a timely manner and require excessive modeling efforts [6]. If early design phases are addressed, then not only is the embedding of the simulation required but also further specific requirements have to be taken into account:

- *MultiLOD*: Scaling and working at different linked levels of detail are important parts of design activities. An important challenge is the feedback of building performance based on such a representation of a digital model of the building with this multiLOD information.
- *Immediate response*: Design sessions need a quick response to steer decisions in the right direction. Designers need to be able to interact with the predictions in real time.
- *Simplified and incomplete information*: The modeling at early design phases is often simplified. Components of the design are not defined in full detail, but simplified descriptions serve as working definitions in the design process. This leads to uncertainty because the energy simulation is missing information. This uncertainty needs to be considered and managed.

To address this demand, we are developing methods that make predictions about the operational energy demand for a building design variant in the early design phase based on machine learning (ML). We use artificial neural networks (ANN), a flexible and universal approach of machine learning. The basis for these methods are component-based ML models [7, 8]. These models allow us to operate with simplified representations that are compatible with early design phases. The models use appropriate decomposition structures, i.e., structures of objects and parameters, that link to the typical representations at these stages. For example, a façade is not represented as a detailed composition of windows and walls but as a window-to-wall ratio parameter.

We link the component-based ML to multiLOD modeling to connect the performance prediction to a digitally enabled design process, which supports interactive design space exploration (DSE) and the performance-based development of solutions. In addition to the ML approach, novel BIM data structures are first required for the multiLOD approach for the early design phases and energy performance of the design. Second, the component-based ML approach must be integrated with the multiLOD approach. The data structures comprise information about the building components in respect to both geometry, e.g., window dimensions, and semantics, e.g., materials and U values. They serve to develop a building design in a way that integrates early performance prediction. The ML component provides instant feedback on energy performance linked to the multiLOD model. Graph rules perform the necessary model transformation from very early LODs to the LOD in which the ML components are applicable. This LOD approach is still more abstract than that of physical building simulation.

1.1 Background

The demand for energy-efficient buildings introduced *building performance simulation (BPS)* into the design process [9, 10]. In early phases of the design process, when important decisions are made, information about design space, which is not just a parametric construct [11, 12], is important. Thus, early-stage BPS searches the design space to guide the design rather than evaluate it. Østergård et al. calls this proactive building simulation [13]. Proactive building simulation represents the interactions that occur between energy analysts, architects, and engineers. 'What if…' questions steer the design search process. However, solutions to the 'what if…' questions may have adverse effects on other performance indicators, such as thermal comfort, daylighting, and visible comfort. Hence, it is important to have a holistic view on performance. Therefore, the main purpose of DSE is to gain this holistic view, as shown by the proposed tools.

Available digital information plays an important role in this context. BIM [1, 2] has been proven as an efficient method to generate, store and manage building information throughout the planning and construction process. There have been many approaches to the implementation of BPS integrated with full-detail BIM data, and these approaches can be differentiated into three types [13]: (1) integrated in the modeling environment, (2) run-time interoperable and (3) file exchange. The specific transformations that are required for these approaches are presented by others, such as Kim and Anderson [6] or van Treeck and Rank [14]. There have also been attempts at semi-automated energy analysis using BIM, in which geometry data are extracted from the BIM model and additional information is integrated by manual input [15]. However, these approaches address full-detail BIM data, which does not exist in the early design phases of buildings. Thus, a workaround is required to add information by defaults and templates, such as those shown by Gervásio et al. [16]. However, the challenge of this workaround is to make realistic explicit assumptions.

Furthermore, researchers have proposed several *rule-based procedures* to add missing information and to formally describe designing. For example, internal space programming algorithms develop the missing information about thermal zones at the early stage of design [17, 18]. ASHRAE 90.1 divides the floor-plan into core and perimeter zones and is a reliable method for performing physical equation-based energy simulations [17, 19]. Moreover, rules based on shape grammars [20], which are a formal approach to generating detailed building components at early design stages, can be developed to perform building performance simulations and predictions [18, 21, 22], which are linked with graph grammars [24, 25]. The idea of these automated processes is to generate a simplified and detailed model of a building that closely resembles the final real building by formally describing the design processes. These approaches are thus a means to add and manage information in a design model.

The *level of development* concept, which formally describes the information accumulation in planning, is relevant in this context. The AIA-LOD definitions and NATSPEC BIM Guide provide definitions for this concept [25, 26]. Several professional and statutory bodies, such as the USA National Institute of Building Sciences [27], the Royal Institute of British Architects (RIBA) [28], Singapore's Building and Construction Authority (BCA) [29] and the Australian National Specification (NATSPEC) [30] have established different levels of developments. There are five BIM model stages during the building lifecycle, which are known as the Conceptual Model, Schematic Design Model, Detailed Design Model, Construction Model and As-built Model. The first three models correspond to the design phase of the building life cycle, while last two are related to the construction and post-construction phases. However, these definitions focus more on information aggregation than on the support design process, with activities such as the crucial practice of scaling. In particular, scaling back and forth is not foreseen in the information aggregation concept of the level of development.

Switching back and forth between scales during the design process, which is called *"scaling"*, is a key activity in design [31, 32]. Abstract design representation is as important as detailed descriptions, and the connection and consistency between the coarse and fine levels of detail are crucial not only for early design phases but also for design in general. To embed such capabilities, novel approaches to BIM covering the information states of early design phases are under development. In the field of tunnel modeling, a multiLOD approach has been established [33, 34]. This approach is based on LOD ideas from GIS [35, 36]; however, these ideas need revision because they focus on reducing information density, whereas the increase in information density is more relevant in design. Furthermore, determining methods for keeping the multiLOD models consistent when information is added at different levels is a key challenge of the field that has been developed for early design phases by the aforementioned research group [5].

In terms of the performance feedback in the early design phases, the *machine learning (ML)* technique for metamodeling energy prediction models has the ability to provide quick performance feedback by capturing all the required BPS interactions with simple model structures and parameters. Multiple studies have shown that ML and other metamodeling techniques can successfully predict the energy performance of buildings [37–41]. As these ML models usually use one network to represent the building, we call them monolithic models. However, the disadvantage of these models is that they are black boxes applicable only to the specific environment they have been trained for. If the structure of the design does not meet the structure the models are trained for, e.g., because new elements and parameters occur in the design, the models cannot be applied. Furthermore, these models do not provide insight into what is occurring in the design, e.g., the heat flows through walls or the temperatures in rooms.

To get more insight into what is occurring in the building design and to transfer and transform ML models more flexibly in building design, we developed a *component-based ML* approach that provides more general components that are transferable to other cases and insights into inter-component behavior, such as heat flows [7, 8]. The component structure in our approach was developed to match the typical ways of describing buildings in design and to match the BIM decompositions. This allows the direct attachment of ML models for energy prediction to BIM, which forms the basis of the developments described in this paper.

1.2 Approach of Component-Based ML in Multi-LOD Modeling

Design and modeling in early design phases typically use low-LOD definitions. These are followed by higher-LOD definitions with more details about the building as soon as the designing proceeds. However, changes occur at all LOD throughout the design processes. Thus, the research was built on the multiLOD approach described in Sect. 2, which is kept consistent by synchronization. In Sect. 3, the component-based ML approach is introduced. ML for multiLOD modeling is prosed, and a method to integrate both approaches is outlined. In Sect. 4, the necessary transformation rules to link different LOD to the ML prediction are developed. Section 5 applies the approach to a test case, shows results for different design scenarios derived from the transformation in early design phases and validates the performance prediction by ML against a detailed simulation in the given design space of the test case.

1.3 Test Case

The researched multiLOD approach and the embedding of ML for energy prediction presented in this paper use the office building of the Tausendpfund construction company as a test case. The building is located in Regensburg in Germany, was built in 2015 and has a net floor area of approximately 1100 m^2 with typical office utilization.

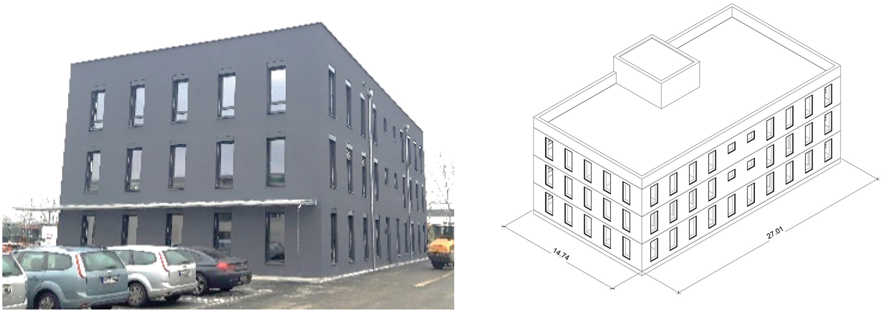


Fig. 1. Photo and Revit model of the Tausendpfund office building used as the test case (courtesy of Tausendpfund GmbH & Co. KG)

2 Multi-LOD Modeling and Energy Prediction

Due to the importance of the scaling described in the background section, the basis of linking the approach to the design process is a multiLOD modeling paradigm. This approach assumes that the design is developed at both coarse low levels of detail and finer high levels of detail (LOD) at the same time. The LOD are adaptive, which means that a designer can modify them as required for a design task. We call this way of modelling adaptive levels of detail (aLOD). Design actions in the early design phases occur at the low levels, whereas the integration of performance prediction and the well-founded evaluation of design options require higher LOD modeling and operations. These are the aLOD definitions that we use:

aLOD 1. The building geometry is defined as a conceptual mass. Only the external surfaces of the building are known. The internal space division is not yet defined.

aLOD 2. The building geometry is defined as a composition of internal spaces defined by components. The thickness, positions and sizes of the components are still not defined.

aLOD 3. The building geometry is defined as a composition of internal spaces with thickness of building components; size and position of openings is also defined.

As shown in Fig. 2, there is first a synchronization between the coarse and fine aLOD in order to propagate changes between the coarser aLOD and the finer aLOD in both directions. Second, a continuous data exchange between the design process and energy analysis process occurs. This first requires a translation of the BIM data, which reflect the design- and planning-related definitions, to the building energy modeling (BEM) data that are required for performance prediction. Second, feedback in terms of performance returns to the BIM environment. At each aLOD, the information becomes more refined as the design progresses. Hence, the assessment of building performance will be more accurate. However, in contrast to the level of development concept, information is propagated back to the low aLOD, allowing the designer to work at these abstract levels to support scaling in the design process.

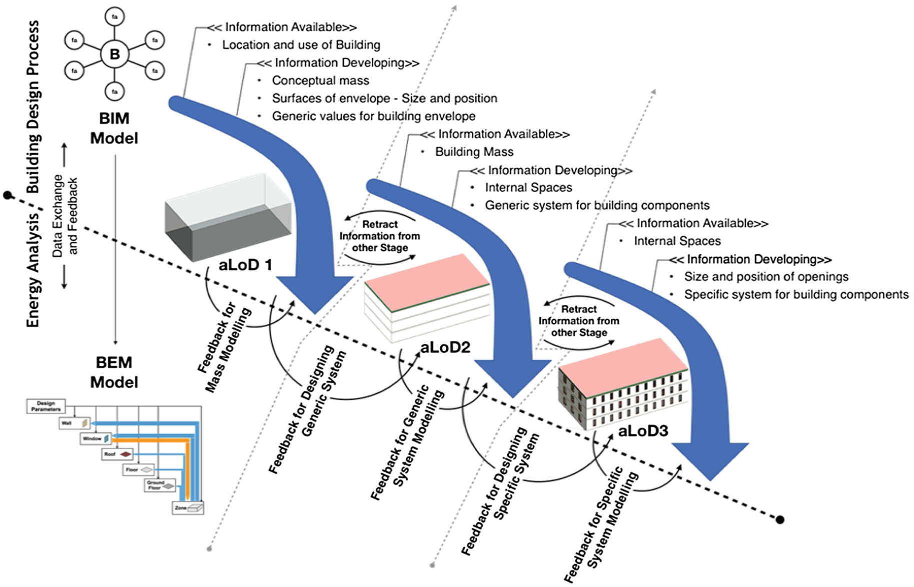

Energy Analysis / Building Design Process
Data Exchange and Feedback
B
fa
BIM Model
BEM Model
Design Parameters
Wall
Window
Roof
Floor
Ground Floor
Zone
<< Information Available>>
• Location and use of Building
<< Information Developing>>
• Conceptual mass
• Surfaces of envelope - Size and position
• Generic values for building envelope
<< Information Available>>
• Building Mass
<< Information Developing>>
• Internal Spaces
• Generic system for building components
<< Information Available>>
• Internal Spaces
<< Information Developing>>
• Size and position of openings
• Specific system for building components
aLoD 1
aLoD2
aLoD3
Retract Information from other Stage
Retract Information from other Stage
Feedback for Mass Modelling
Feedback for Designing Generic System
Feedback for Generic System Modelling
Feedback for Designing Specific System
Feedback for Specific System Modelling

Fig. 2. Interaction between building design and energy analysis process.

2.1 Data Structures for MultiLOD

To capture information for operational energy prediction in the BEM, we propose the data structure described in Fig. 3. This structure represents a block definition diagram (bbd) according to the systems modeling language (SysML, [42]) for a BEM.

As shown here, a building is composed of zones. A zone can be defined as an element with a similar occupancy or usage throughout the building, such as an office, technical services, and a corridor. A zone has internal heat sources, such as occupants, lighting, and equipment, as well as equal internal climate conditions. Moreover, a zone has a spatial enclosure to uniquely define it in a three-dimensional system. A zone interacts thermally with the environment through a spatial enclosure. A spatial enclosure is defined by a number of wall, floor and roof elements. Any heat exchange with the environment or another zone is calculated based on the properties of these elements. The wall, floor and roof elements are considered specific blocks of *"ThermalFace"*, as

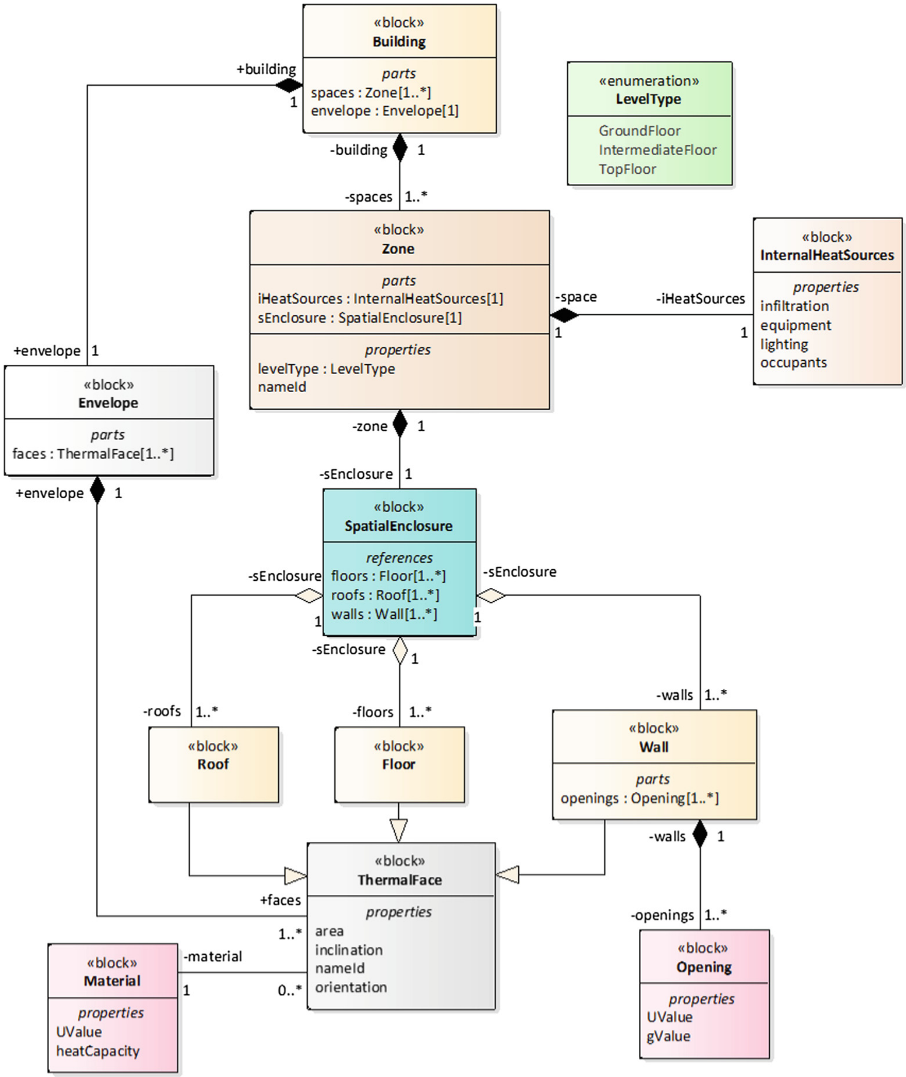


Fig. 3. Data structure for information exchange between BIM and energy prediction model.

they share many properties and behaviors in terms of heat exchange. The block *"Wall"* is composed of openings to represent the doors and windows hosted in any wall.

The data structure includes the parameters that are required for the energy performance prediction, such as dimensions, U values, g values, and heat capacities. Furthermore, it is prepared to attach ML models to the objects in order to facilitate energy prediction, as the structure of the design is put together as described in the following sections.

3 Component-Based Machine Learning for MultiLOD Modeling

Physical simulations of operational energy, among the most important simulations in building design, require full-detail modeling with a high modeling effort and information demand and long computation times for the design processes. Therefore, a component-based ML approach has been developed that is tailored to the design process and the available information, such as typical objects and parameters. The following subsection will describe the approach, while Subsect. 3.2 will outline how the component-based ML approach is integrated into multiLOD modeling.

3.1 ML Components and Their Training

The paradigm of component-based ML is that the typical design configuration is broken down to a set of parametric ML models that are plugged together according to the available information in the design process to predict the performance of the building, e.g., the primary energy efficiency. Figure 4 shows the decomposition that has been developed for this purpose. This structure links to the data structure shown in the previous section and utilizes the information for a ML-based prediction of the energy performance of the building. With the design parameters, the first level heat flow predictions are performed by ML components for the wall, floor or roof component. A machine learning model aggregates the heat exchange between a zone and the environment via these components. Geometrical and material parameters, along with zone properties such as air changes and floor level, are used to make the heat flow prediction. Based on this heat flow prediction, the zone component aggregates the information by means of another ML model to predict the heat and cold demands of a zone. These coupled ML layers in the components form an engineering-driven deep learning approach, as described previously [43, 44].

For the development of the ML components, training data from a simulation model that cover the relevant conditions for the components are used. In this case, the word relevant means that the design parameter ranges are sufficiently covered in the training data and that the ranges of the internal flow parameters, such as heat flow, meet those of the design cases. It is crucial that the model covers the relevant conditions; otherwise, the use of the ML model will leave the validated design space predictions at risk for high inaccuracies. The model for generating training data is a parametric simulation model built by Energy Plus software, as shown in Fig. 5, left. The training data generation occurs at an LOD corresponding to aLOD 3, which is the LOD of the ML components.

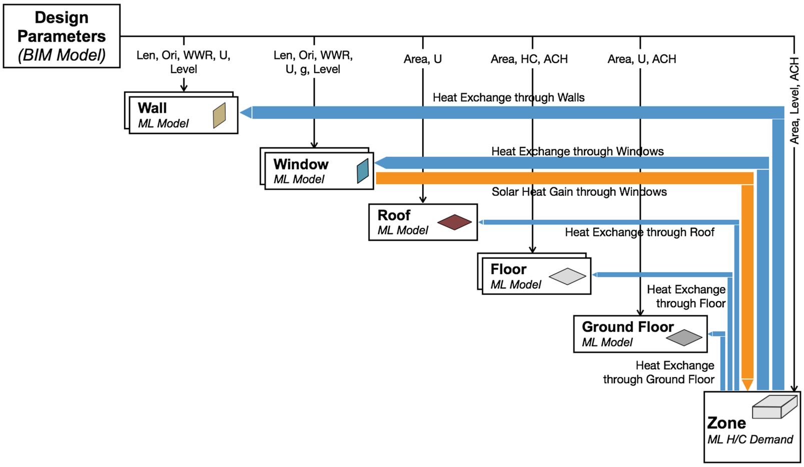
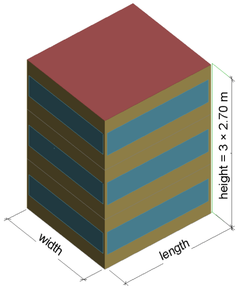

Design Parameters (BIM Model)
Len, Ori, WWR, U, Level
Len, Ori, WWR, U, g, Level
Area, U
Area, HC, ACH
Area, U, ACH
Area, Level, ACH
Wall ML Model
Heat Exchange through Walls
Window ML Model
Heat Exchange through Windows
Solar Heat Gain through Windows
Roof ML Model
Heat Exchange through Roof
Floor ML Model
Heat Exchange through Floor
Ground Floor ML Model
Heat Exchange through Ground Floor
Zone ML H/C Demand

Fig. 4. System flow diagram for decomposition for energy prediction



		Unit	min	max
Length		m	10	80
Width		m	10	80
Window-to-wall ratio (WWR)	S		0%	95%
	N		0%	95%
	E		0%	95%
	W		0%	95%
ORI			-180°	180°
U value walls		W/m^{2}K	0.411	0.776
U values windows		W/m^{2}K	0.5	2
U value ground floor		W/m^{2}K	0.411	0.864
Heat capacity floors		J/kgK	900	1200
U value roof		W/m^{2}K	0.191	0.434
g value windows			0.5	0.95
Air change rate		h^{-1}	0.2	1

Fig. 5. Ranges of the training data to generate the machine learning components.

On the right side of Fig. 5, the ranges for which the training model produced data following a Latin hyper-cube (LHS) experimental design with 400 runs are shown. A typical cross validation strategy splitting the dataset into training data (70%), validation data (15%) and test data (15%) is used to train the sigmoid function artificial neural network (ANN) with 5 to 10 neurons, either by the Levenberg-Marquardt method or by the Bayesian regularization method in MatLab. A set of ML components results from this process. Artificial test cases have shown that predictions with no more than 4%

error for heating and cooling are possible; details are described in a previous work [8]. Prediction times can be significantly reduced by the approach, e.g., from 1145 s to 0.9 s, for approximately 200 design configurations, which allows design space exploration to be truly integrated in the design process [45].

3.2 Integration of Component-Based ML and MultiLOD Modeling

The definitions of low aLOD are based on abstract definitions with parameters tailored to important design properties, such as the window-to-wall ratio (WWR), for urban massing at the lowest LOD (Fig. 6, left). The physical building performance simulation (BPS) requires the full level of detail (Fig. 6, lower sequence). This necessitates all relevant details of the building to be modelled to perform the simulation, which is related to the previously mentioned high effort and computation time. This effort is not appropriate for early design phases. Thus, alternatively, simplified physical simulation models can support the design process. However, simplified approaches also require manual definitions of the simplifications and dynamic BPS models.

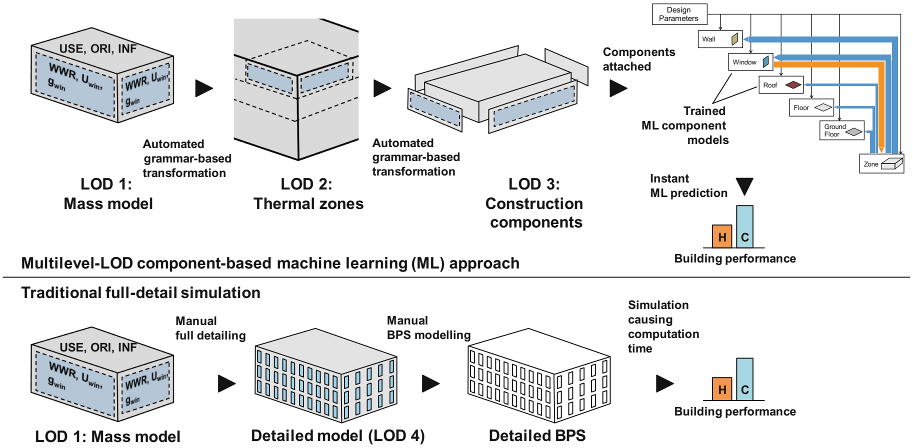

USE, ORI, INF
WWR, U_{win}, g_{win}
WWR, U_{win}, g_{win}
Automated grammar-based transformation
LOD 1: Mass model
LOD 2: Thermal zones
Automated grammar-based transformation
LOD 3: Construction components
Components attached
Design Parameters
Wall
Window
Roof
Floor
Ground Floor
Zone
Trained ML component models
Instant ML prediction
H
C
Building performance
Multilevel-LOD component-based machine learning (ML) approach
Traditional full-detail simulation
USE, ORI, INF
WWR, U_{win}, g_{win}
WWR, U_{win}, g_{win}
Manual full detailing
Manual BPS modelling
Simulation causing computation time
LOD 1: Mass model
Detailed model (LOD 4)
Detailed BPS
H
C
Building performance

Fig. 6. Structure of the multiLOD machine learning (ML) approach compared to the traditional building performance simulation (BPS).

Due to the reasons given above, the described ML component approach was developed. The components are tailored to relevant design parameters and key performance indicators to guide the design process. ML models act as metamodels or surrogate models and represent important performance aspects linked to early-design-stage models (Fig. 6, top right).

For this purpose, it is necessary to couple the component-based ML model to the different LOD that comprise different objects with different information densities. Therefore, the linking of the ML components to the LOD is required. Two different approaches are possible:

1. An approach that develops ML components for every LOD. As soon as the LOD is switched, another ML component set is used, and all predictions start from scratch.
2. An approach that develops ML components at a suitable LOD for the energy prediction. In this case, suitable means that the information with the right detail structure in terms of the components and parameters is foreseen in the ML decomposition. Lower or higher LOD need to be transformed to the ML LOD.

We decided to follow the second approach. The utilization of one decomposition level to develop the ML component instead of developing ML components for every LOD has the following advantages:

- It allows for an adaptive and dynamic definition of information at different LOD with the continuous observation of the reactions in the configuration. This continuous model development also includes the possibility to only partially update the ML model in case of changes. The adaptiveness of LOD would cause problems in the first approach, as components or parameters might be missing in the ML set.
- In contrast to monolithic ML models (a low-LOD ML model following the first approach would be a monolithic model), the internal information on system flows, e.g., heat flows between components, is always available, even at the lowest LOD. This provides designers with information to steer the design process.
- Inconsistencies between ML models at different LOD are avoided. When ML models are provided for each LOD separately, switching between LOD will probably lead to inconsistencies between the ML models. For instance, the results predicted at one LOD might differ from the results predicted at a higher or lower LOD in a way that is not understandable for the designer. Transforming low LOD to higher LOD makes deviations understandable because transformed low-LOD models can be compared to high LOD models.

The linking of different LOD to one component-based ML requires transformations of multiLOD modeling to ML. The sequence at the top of Fig. 6 illustrates this transformation process. An approach for these transformations is developed in the next section.

4 Transformation in MultiLOD Modeling

The ML components in the previous section correspond to aLOD 3 according to the definition in Sect. 2. To attach the lower LOD to the components, we propose transformation rules that form a design grammar. This grammar describes the detailing process and adds components and structures to allow the ML-based energy performance prediction, including with low-LOD information.

As evident in Fig. 2, the information about the construction components is not present at the beginning of the design process i.e., aLOD1 and aLOD2, so there is a need to generate information about these components to setup a BEM model. We have developed design grammar rules as represented in Fig. 7 to convert an early stage BIM model for information extraction. A design grammar in this context is a combination of a shape grammar dealing with geometry and a graph grammar dealing with engineering

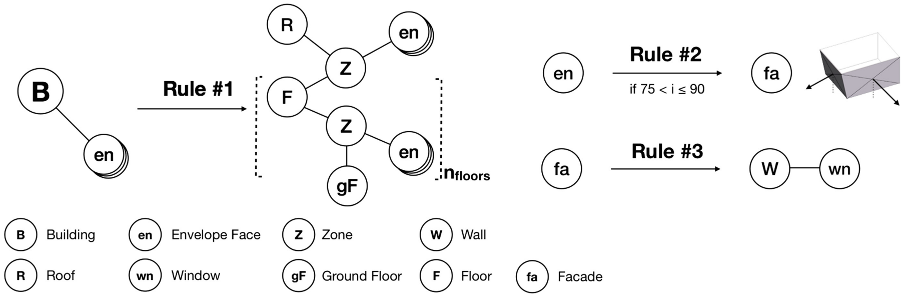


Fig. 7. Design grammar rules to generate construction components.

aspects. The rules of this design grammar are applied to generate geometrical and parametrical construction components in the BIM model. These components are needed for the ML prediction. The material information is associated with the components based on user inputs or default assumptions.

The grammar to transform low aLOD to the ML aLOD comprises three rules:

Rule #1: This rule is applied at aLOD 1 to generate *"Floors"* and a *"Zone"* corresponding to each floor. The face of the building envelope that is in contact with the ground is transformed to the *"Ground Floor"* element. The full height of mass is divided by the number of floors based on the total floor area required, the floor height and the building mass footprint. At each level, a *"Floor"* is created with a corresponding *"Zone"* at that level. The envelope face that is facing upwards is converted to the *"Roof"* component.

Rule #2: This rule is applied to transform the envelope faces to *"Façade"* components. All the envelope faces that are inclined between 75° and 90° are converted to *"Façade"*.

Rule #3: This rule is applied to transform the *"Façade"* components to the *"Wall"* and *"Window"* components. The user enters the value of the window-to-wall ratio for each direction to generate the window components in the center of the wall.

The developed shape grammar rules follow the typical design process, as *Rule #1* and *Rule #2* transform the building representation to a form similar to the aLOD 2 representation, and *Rule #3* transforms the aLOD 2 representation to aLOD 3.

Figure 8 describes the transformation of an early design state of the Tausendpfund building model to a detailed construction component model for energy performance prediction. The design grammar rules (Fig. 7) serve this purpose. The design process starts with "aLOD 1: Start"; after implementing *Rule #1* and *Rule #2*, the building representation is transformed to a building representation corresponding to aLOD 2, i.e., the internal space divisions, and *Rule #3* transforms the aLOD 2 model to a representation that corresponds to aLOD 3, i.e., a detailed model.

In Fig. 8, A (i) shows a graph-based representation of a very early BIM model, and B (i) shows its visualization at the start of aLOD 1. After applying *Rule #1*, the initial aLOD 1 is transformed to A (ii). *"Floors"* and *"Zones"* are created at each floor, and a roof component is created at the top. B (ii) visualizes the BIM model at aLOD 1: Step1. After this, *Rule #2* is applied, which transforms the envelope faces to *"Façade"* components. The transformation is expressed in A (iii) aLOD: Step 2. B (iii) is the BIM model visualization at the corresponding step. Finally, *"Rule #3"* is implemented to generate *"Wall"* and *"Window"* components out of the *"Façade"* components. A (iv) and B (iv) are the graph representation and the BIM model visualization, respectively, after the implementation of *Rule #3*.

A (iii) also corresponds to the aLOD 2 representation, i.e., the building representation expressed in A (iii) is the same as the aLOD2 representation. The model at aLOD 2 can be transformed for information extraction by implementing *Rule #3*. Once the construction components are generated in the BIM model, thermal properties are associated with the construction components according to user input or default assumptions.

It is important to note that the transformation rules add information by detailing the design configuration. As this information might not be defined and might include unchecked assumptions. Therefore, if this information is not provided by the designer or engineer, it is required to inform him/her about the assumptions and the connected range of variation in the prediction. For instance, the different number of stories, as shown in Fig. 8, or added parameters, such as U values and thermal capacities, form variants that require the addition of more information in the design progress. Unless defined finally, these options remain open to variations and need to be checked in the evaluation of an early design phase configuration.

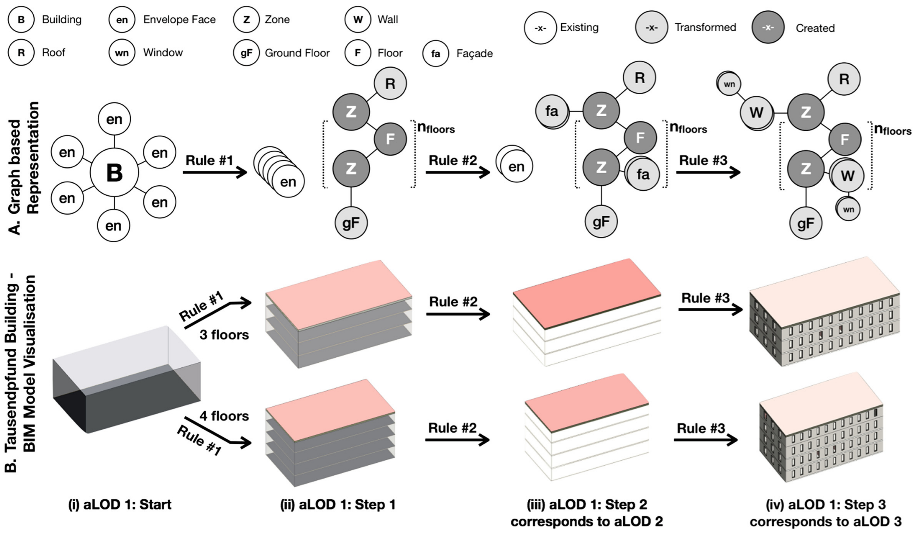


Fig. 8. Transformation process of the Tausendpfund case building from the lowest aLOD to ML aLOD.

5 Validation by Means of the Test Case

To validate the component-based ML approach, it was tested by means of the Tausendpfund test case (Sect. 1.3 and Fig. 1). As the starting point, a model of the building mass at an early design phase corresponding to aLOD 1 with some energy-relevant design properties was been assumed (Fig. 8, left). The transformation procedure described in Sect. 4 (Fig. 8) led to a set of possible solutions for the building case at the ML LOD (aLOD 3). Two different designs types are possible by the transformation: a three-story and a four-story design. For validation, the parametric physical simulation model in Energy Plus was configured to the same design variants as the case and compared to the ML predictions. This was performed for 60 different configurations in the range of the case's design configuration, as shown in Table 1. Width, depth and height have a limited range due to the required floor area in the architectural program and the conditions of the site, which allow three or four storeys, leading to two height options of 8.10 or 10.80 meters with corresponding widths and depths. For the window-to-wall ratios, more variation is possible. Thus, the full range of options was tested. The orientation was defined by the site and was not subject to variation. The U values of the envelope and g values of the windows, as well as the air change rate, depend on the façade construction and operation of the building; thus, a realistic range was considered.

Table 1. Configurations for design space exploration

Parameter	Unit	Built	Test range	Comments
Width	m	26.03	26.03, 22.78	Depending on the number of storeys generated by the transformation rules
Depth	m	14.03	14.03, 12.15	
Height	m	8.10	8.10, 10.80	
Window-to-wall ratio	S	27%	1…95%	Percentage of window area in the façade
	N	27%	1…95%	
	E	26%	1…95%	
	W	30%	1…95%	
Orientation		18°	18°	Kept constant as given by the side
Wall U value	W/m^{2}K	0.4	0.4…0.78	
Window U value	W/m^{2}K	0.87	0.5…2	
Ground floor U value	W/m^{2}K	0.4	0.4…0.86	
Intermediate floor heat capacity	J/kgK	1150	900…1200	
Roof U value	W/m^{2}K	0.4	0.19…0.43	
Window g value		26%	50…95%	Slightly adapted to the validated ML model range
ACH	h^{-1}	1.3	0.2…1	

With these parameters, a test series of 60 runs exploring the available design space was performed. The component-based ML was applied to the multiLOD model approach to predict the energy demand of the building within this design space. This means structures are derived via the transformation rules, parameters are added to the

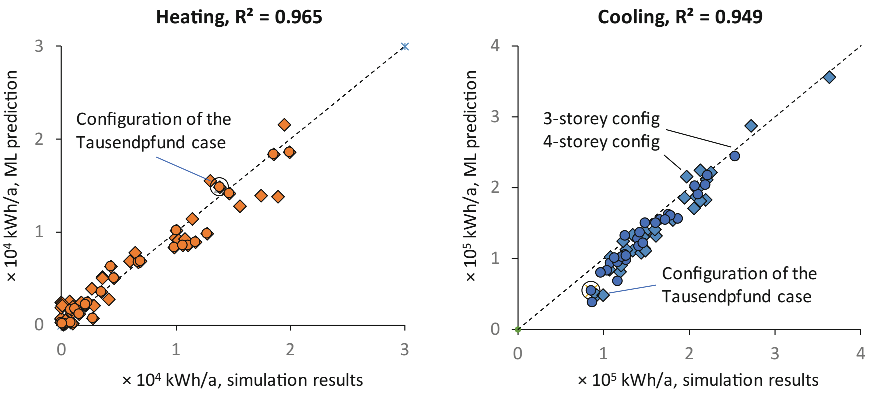


Fig. 9. Validation of ML prediction against parametric simulation for the test case.

ML components and energy performance is predicted. This occurred in an automated process. As references, two detailed parametric simulation models were developed in Energy Plus, one with the three storeys and one with four storeys.

The comparison shown in Fig. 9 proved that the prediction within the design space of the case is quite reliable. The coefficient of determination R^2 is 0.965 for the heating energy demand and 0.949 for the cooling energy demand. The maximum error for heating is 5.1 × 10^3 kWh/a, which is 25% of the maximum. For cooling, the maximum error is 5.0 × 10^4 kWh/a, which is 14% of the maximum. The average errors are 3.8% for heating and 3.5% for cooling, which are sufficient for early design phases.

6 Discussion

The integration of a component-based ML approach in multiLOD modeling has been successfully shown. The basis for this integration is a transformative design grammar. The accuracy, which is sufficient for early design phases, will be improved by further research examining decomposition approaches and parameters and ML models in more detail, such as those described in our previous works [43, 44].

Decomposition is not only a question of model accuracy; we would also like to emphasize that the use of ML components, their integration with existing BIM structures and the multiLOD approach, which are founded on long-term experience in design and engineering, are crucial for selected decomposition strategies. The aim of the ML components is not only to provide a good prediction but also to be well integrated into modeling processes of design and engineering. In an ideal situation, ML components can be attached one-to-one to the modeling objects.

Furthermore, important aspects of decomposition are the system flows between components. These flows provide designers and engineers with insight into why a design configuration performs well or poorly with respect to the performance targets. The decomposition structure, with its components and parameters, determines which system flow in-between the components can be observed. Therefore, the decomposition has to reflect the planned use of components and fit well to the engineering tasks taking place. The necessary components and parameters need to be present, otherwise the component set is not able to support the engineering task.

Until now, one stringent decomposition level of ML that is interconnected with the multiLOD approach by model transformations has been developed. As mentioned in Sect. 3, this study made a conscious decision to avoid switching between ML models, which might cause gaps in the transitions and further inconsistencies. However, this decision comes at the cost of requiring the automated detailing of the low LOD, which was shown in Sect. 4. Nonetheless, we deem the advantages of this approach to outweigh its disadvantages. Further examinations of test cases will provide more practical information about this decision.

The generalization of the developed ML models has successfully been tested for rectangular non-box designs [8]. We also expect that the approach works for non-rectangular designs, given that key features, such as the ratio of the façade area to volume, are in the range of the training data. Additional model changes, such as different occupation patterns or heating controls, also seem possible. However, both are subject to further research.

The detailing by a design grammar is demonstrated in Sect. 4. The defined transformation rules formally describe a possible detailing of the design. It is important to note that they do not replace the design process and prescribe how to develop the design. In a user interface and in the model, they do not need to be shown to the user to suggest design solutions. Rather, the user should note the fact that the possible variations and information to be added lead to a possible performance range of the design. Furthermore, an analysis that we will perform in future research will highlight the impact made by the addition of the information. The aim is to motivate the user to think about making decisions with high impacts first to quickly gain knowledge about the solution performance without progressing too much in detailing a building design. This reduces efforts at the begin of the design process and allows the examination of more variable results for better designs.

7 Conclusions

An approach where multiLOD integrates component-based ML has been described in this work. The adaptive LOD support early phases of building design by enabling designers and engineers to perform the important activity of scaling with embedded performance prediction. The consistent LOD allow the development and understanding of a design at different degrees of abstraction or details, which correspond to the core design activity of “scaling”.

The integration of trained ML components provides a quick prediction of the energy performance embedded in the design activity. It has been shown that ML components attached to BIM objects to represent performance provide, in their composition, a very fast prediction of performance with a sufficient accuracy to steer the design process.

Furthermore, transformation rules have been established to allow the flexible integration of early-stage BIM data with the component-based ML approach. Given the very high prediction speed of ML, which reduces computation time to roughly a thousandth of time, the automated transformation provides a good approach to detail low LOD to the ML component LOD with many variations. Such variations are caused by missing information. The approach allows design space exploration to determine the range of performance. This includes access to the variations caused by missing information.

This approach gives designers a more realistic picture of the performance. Further analysis of the design space allows the determination of which decisions make the greatest impacts and should be prioritized. In this way, the multiLOD modeling approach combined with component-based ML provides energy performance predictions to guide the early design processes.

Acknowledgements. The research presented in this paper was funded by a KU Leuven starting grant StG/14/020 and by the German Research Foundation (DFG) in the Researcher Unit 2363 “Evaluation of building design variants in early phases using adaptive levels of detail” in Subproject 4 “System-based Simulation of Energy Flows”. The multiLOD approach was developed in discussion with the members of the research group (DFG Researcher Unit).

References

1. Eastman, C., Teicholz, P., Sacks, R., Liston, K.: BIM Handbook. Wiley, Hoboken (2011)
2. Borrmann, A., König, M., Koch, C., Beetz, J.: Building Information Modeling - Technologische Grundlagen und industrielle Praxis. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-658-05606-3
3. Sawhney, A., Maheswari, J.U.: Design Coordination Using Cloud-based Smart Building Element Models. Int. J. Comput. Inf. Syst. Ind. Manag. Appl. **5**, 445–453 (2013)
4. BuildingSMART: Industry Foundation Classes IFC2x4 (2013). http://www.buildingsmart-tech.org/specifications/ifc-releases/ifc4-release/
5. König, M., Borrmann, A., Geyer, P., Schneider, P., Lang, W., Petzold, F., Schellenbach-Held, M.: Evaluation of building design variants in early phases on the basis of adaptive detailing strategies and system-based simulation of energy flows for such models. In: Presented at the Beyond BIM Workshop, Ghent (2017)
6. Kim, H., Anderson, K.: Energy modeling system using building information modeling open standards. J. Comput. Civ. Engi. **27**, 203–211 (2013)
7. Singaravel, S., Geyer, P., Suykens, J.: Component-based Machine Learning Modelling Approach For Design Stage Building Energy Prediction. In: IBPSA 2017 (2017)
8. Geyer, P., Singaravel, S.: Component-based building performance prediction using systems engineering and machine learning. Appl. Energy (2018, submitted)
9. Schlueter, A., Thesseling, F.: Building information model based energy/exergy performance assessment in early design stages. Autom. Constr. **18**, 153–163 (2009)
10. Clarke, J., Hensen, J.: Integrated building performance simulation: progress, prospects and requirements. Build. Environ. **91**, 294–306 (2015)

11. Gero, J.S., Kumar, B.: Expanding design spaces through new design variables. Des. Stud. **14**, 210–221 (1993)
12. Gane, V., Haymaker, J.: Design scenarios: enabling transparent parametric design spaces. Adv. Eng. Inform. **26**, 618–640 (2012)
13. Østergård, T., Jensen, R.L., Maagaard, S.E.: Building simulations supporting decision making in early design – a review. Renew. Sustain. Energy Rev. **61**, 187–201 (2016)
14. van Treeck, C., Rank, E.: Dimensional reduction of 3D building models using graph theory and its application in building energy simulation. Eng. Comput. **23**, 109–122 (2007)
15. Ahn, K.U., Kim, Y.J., Park, C.S., Kim, I., Lee, K.: BIM interface for full vs. semi-automated building energy simulation. Energy Build. **68**, 671–678 (2014)
16. Gervásio, H., Santos, P., Martins, R., da Silva, L.S.: A macro-component approach for the assessment of building sustainability in early stages of design. Build. Environ. **73**, 256–270 (2014)
17. Dogan, T., Reinhart, C., Michalatos, P.: Autozoner: an algorithm for automatic thermal zoning of buildings with unknown interior space definitions. J. Build. Perform. Simul. **9**, 176–189 (2016)
18. Granadeiro, V., Duarte, J.P., Palensky, P.: Building envelope shape design using a shape grammar-based parametric design system integrating energy simulation. In: IEEE Africon 2011, pp. 1–6. IEEE, Livingstone (2011)
19. ASHRAE: Standard 90.1–2013. Energy Standard for Buildings Except Low-Rise Residential Buildings, Atlanta, GA (2013)
20. Stiny, G.: Shape: Talking About Seeing and Doing. MIT Press, Cambridge (2006)
21. Mitchell, W.J., Liggett, R.S., Pollalis, S.N., Tan, M.: Integrating shape grammars and design analysis. In: Computer Aided Architectural Design Futures: Education, Research, Applications, proceedings of CAAD Futures 1991. pp. 17–32. CAAD Futures, Zürich (1991)
22. Geyer, P.: Multidisciplinary grammars supporting design optimization of buildings. Res. Eng. Des. **18**, 197–216 (2008)
23. Schmidt, L.C., Shetty, H., Chase, S.C.: A graph grammar approach for structure synthesis of mechanisms. J. Mech. Des. **122**, 371–376 (2000)
24. Helms, B., Shea, K.: Computational synthesis of product architectures based on object-oriented graph grammars. J. Mech. Des. **134**, 021008-14 (2012)
25. BIMForum: Level of Development Specification (2016)
26. NATSPEC Construction Information: NATSPEC National BIM Guide (2011)
27. National Institute of Building Sciences buildingSMART alliance: National BIM Standard - United States $^{\mathrm{TM}}$ Version 2 (2015)
28. National Building Specifications: NBS BIM Toolkit. https://toolkit.thenbs.com/support
29. Building and Construction Authority: Singapore BIM Guide - Version 2.0., Singapore (2013)
30. NATSPEC BIM - NATSPEC National BIM Guide. https://bim.natspec.org/documents/natspec-national-bim-guide
31. Yaneva, A.: Scaling up and down: extraction trials in architectural design. Soc. Stud. Sci. **35**, 867–894 (2005)
32. Ammon, S.: Why designing is not experimenting: design methods, epistemic praxis and strategies of knowledge acquisition in architecture. Philos. Technol. **30**, 495–520 (2017)
33. Borrmann, A., Kolbe, T., Donaubauer, A., Steuer, H., Jubierre, J.R.: Transferring multi-scale approaches from 3D city modeling to IFC-based tunnel modeling. In: Proceedings of the 3DGeoInfo, pp. 27–29 (2013)
34. Borrmann, A., Jubierre, J.R.: A multi-scale tunnel product model providing coherent geometry and semantics. In: Proceedings of the 2013 ASCE International Workshop on Computing in Civil Engineering, pp. 1–8 (2013)

35. Meng, L., Forberg, A.: 3D building generalisation. Gen. Geogr. Inf. Cartogr. Model. Appl. Elsevier. (2007)
36. Glander, T., Döllner, J.: Abstract representations for interactive visualization of virtual 3D city models. Comput. Environ. Urban Syst. **33**, 375–387 (2009)
37. Eisenhower, B., O'Neill, Z., Narayanan, S., Fonoberov, V.: A methodology for meta-model based optimization in building energy models. Energy Build. **47**, 292–301 (2012)
38. Magoulès, F., Zhao, H.: Data Mining and Machine Learning in Building Energy Analysis. Wiley Online Library (2016)
39. Van Gelder, L., Das, P., Janssen, H., Roels, S.: Comparative study of metamodelling techniques in building energy simulation: guidelines for practitioners. Simul. Model. Pract. Theory **49**, 245–257 (2014)
40. Stavrakakis, G., Zervas, P., Sarimveis, H., Markatos, N.: Optimization of window-openings design for thermal comfort in naturally ventilated buildings. Appl. Math. Model. **36**, 193–211 (2012)
41. Cheng, M.-Y., Cao, M.-T.: Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. Appl. Soft Comput. **22**, 178–188 (2014)
42. Object Management Group: Systems Modeling Language, Specifications Version 1.5. http://www.omg.org/spec/SysML/1.5/
43. Singaravel, S., Geyer, P., Suykens, J.: Deep neural network architectures for component-based machine learning model in building energy predictions. Presented at the eg-ice Workshop 2017 (2017)
44. Singaravel, S., Geyer, P., Suykens, J.: Deep learning neural networks architectures and methods: building design energy prediction by component-based models. Adv. Eng. Inform. (2018, submitted)

Detect Relationship Between Urban Housing Development and Urban Heat Island Dynamic in Hyper-density Hong Kong by Integrating GIS and RS Techniques

Jin Yeu Tsou[1,2](✉), Xiang Li[2], Katerina Tsou[3], Jiahui He[4], and Dongxu Pan[5]

[1] Center for Housing Innovations, The Chinese University of Hong Kong, Shatin, Hong Kong
jinyeutsou@cuhk.edu.hk

[2] School of Architecture, The Chinese University of Hong Kong, Shatin, Hong Kong

[3] Department of Civil and Environmental Engineering, University of California, Berkeley, USA

[4] MSc. Programme in Advanced Environmental Planning Technologies, The Chinese University of Hong Kong, Shatin, Hong Kong

[5] MSSc. Programme in Housing Studies, The Chinese University of Hong Kong, Shatin, Hong Kong

Abstract. Urbanization and aging of society are two converging trends of current demographic changes. The intensified human activities associate with the formation and dynamic of Urban Heat Island (UHI) which is harmful to health. Looking into the correlation between UHI effect and the land surface coverage of everyday spaces is curtail and significant. Aided by spatial analysis and spatial statistic functions based on Geographical Information System (GIS) and data extraction and processing methods enabled by the Remote Sensing (RS) platform, this paper detected the distribution of land surface temperature and traced its changes alongside dynamics of land surface coverage in the selected typical areas of Hong Kong. Findings of this paper will be significant for Hong Kong planners, architects and housing officials to consider when deciding on the next step of Hong Kong urban planning and housing development.

Keywords: Urban housing · High-density · Hong Kong
Geographical Information System · Remote Sensing

1 Introduction

1.1 High-Density Housing Development in Hong Kong

The speed of urbanization, especially in Asia mega cities, is incredible. Hong Kong, already well known as a high-density city, is coming across the more serious challenge of increasing population and decreasing land resources. Since 1841, Hong Kong has been successfully developed into an international financial center, super high-density urbanized area, through efficient land resource management as well as mixed and intensified development mechanism. During the 1950s, the baby boom and influx of massive immigrants boost the tremendous increase of population, which laid a solid foundation

© Springer International Publishing AG, part of Springer Nature 2018

I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 535–553, 2018.
https://doi.org/10.1007/978-3-319-91635-4_28

for the expansion of labor intensive industries, such as plastic flower and toy manufacturing. To obtain working opportunities, most of the population settled in the Kowloon, which is one of the earliest-exploited areas and social-economic centers at the time. Squatter hubs sprawled and clustered. The crowded and poor quality of living environment brought hygiene and fire safety hazards. Though attentions to these hazards were paid by the government, the necessity to develop public housing scheme was not brought to forefront since it was literally a big challenge for housing supply to cope with the rapid population growth. Until the end of 1953, a big fire in Shek Kip Mei left nearly 53000 people homeless overnight, whilst became the catalyst of the birth of Hong Kong Public Housing scheme and pushed forward the people-oriented housing approach.

Housing in Hong Kong is mainly comprised of 3 types, i.e. public rental housing, Home Ownership Scheme (government subsidized housing) and private housing. The census revealed that there are about 7.3895 million of people living in Hong Kong by mid of 2017. Only 3.7% of the total 1111 sq.km land is available to accommodate the citizens' housing demand, whilst 1% of land is used for public housing that provide home for about 30% of total Hong Kong populations (Fig. 1). The housing strategy in Hong Kong made a big success to stabilize the social process and economic flourish of the city in the long run.

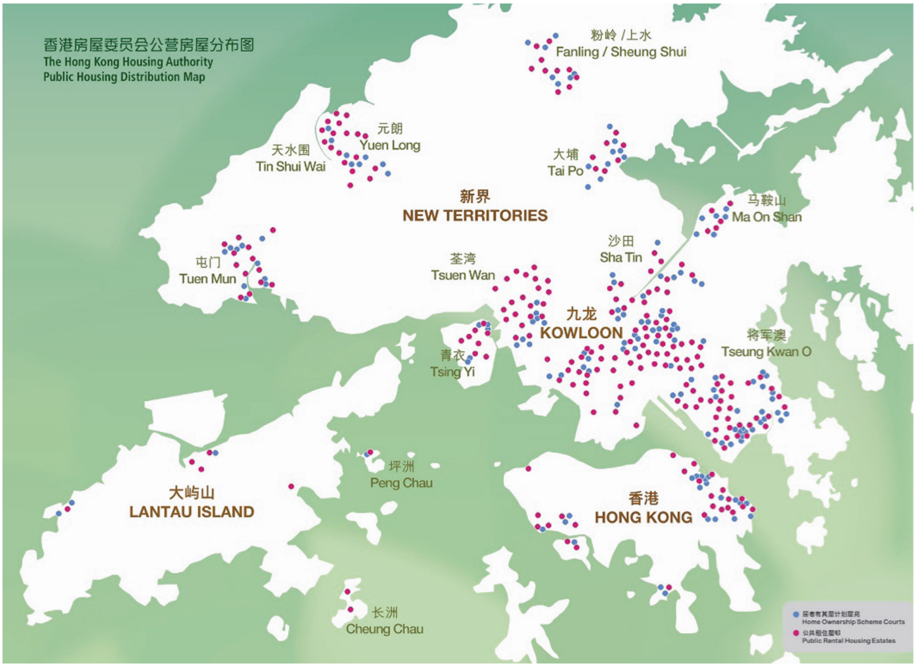


Fig. 1. Distribution of Hong Kong public housings (Resource: *The Hong Kong Housing Authority Public Housing Portfolio.* Hong Kong Housing Authority, Mar 2017)

Hong Kong housing long maintained high plot ratio and proximity of buildings. The overall density of living is 6780 people per square kilometer in the mid-2016 according to census of Hong Kong. Due to the extremely high urban density, the "wall building" is commonly seen in urbanized areas of Hong Kong, which blocks ventilation and congregates absorbed solar heat. In 2003, Hong Kong suffered from SARS with 1755 people infected and 299 people unfortunately lost their lives. Amoy Garden, a residential estate in Kwun Tong was the severely affected area. The poor wind and lighting environment of residential units caused by the extremely high building density was reckoned as the leading reason. The SARS wakes up Hong Kong that the context of urban living environment can have consequential influence on the quality of live. Afterwards, the Air Ventilation Assessment System was established by the Hong Kong Planning Department to guide later building and planning projects in micro-climate studies for sustainable living environmental design.

Under the housing pressure of steady population growth (Table 1), Hong Kong has kept the efforts to innovate housing typology to satisfy the divergent social and environmental contexts in different stages during the past decades. However, the super high urban density is still a big challenge to cope with a number of urban environmental issues, such as Urban Heat Island effect.

Table 1. Domestic households by type of housing (Resource: Hong Kong Annual Digest of Statistics)

Type of housing	Number of new domestic households developed (Million)					
	2006	2011	2012	2013	2014	2015
Public rental housing	6.746	7.087	7.291	7.341	7.412	7.596
Subsidized home ownership housing	3.628	3.722	3.73	3.711	3.751	3.75
Private permanent housing	11.647	12.597	12.672	12.864	12.992	13.215
Temporary housing	0.187	0.187	0.169	0.158	0.169	0.15
Total	22.209	23.593	23.862	24.073	24.324	24.711

1.2 Urban Heat Island Effect

Urban Heat Island (UHI) is a term to describe the urban climatic phenomenon that the ambient temperature of dense urban core is higher than that of the rural environment and the nocturnal UHI effect is more apparent than daytime UHI (Kyriakodis and Santamouris 2017). The formation and dynamics of the UHI are manifested to be caused by positive thermal balance of cities due to overdoing heat absorption of the building facades and urban infrastructures (Santamouris 2013). Both daytime and nocturnal UHI are influenced by a variety of factors which could be mainly summarized as canopy condition and land-cover. In terms of canopy condition, the influential factors of UHI effect include sky view factor, surface albedo and surface roughness etc., whilst land-cover manipulates the UHI dynamics through changing of proportion of impervious layers, i.e. the concrete or asphalt paving, against pervious layers, i.e. vegetation cover and water body.

Due to the complexity of urbanization in recent days, the cause of UHI formation and dynamic are multi-faceted. The mechanisms of influence are varied based on the type of construction activity, which is likely to result into diversified pattern of population flow and cluster. The traditional single-aspect data-driven studies could not demonstrate holistically the contributions made to the UHI formation or dynamics. The super high density of urban Hong Kong can't avoid the side-effects from UHI resulted by the overcrowded buildings and overwhelming human activities. Residential buildings, as the fundamental urban infrastructure, are built aligning with the population growth. Getting clear about the correlation between UHI dynamic and housing or community development is of great importance to ameliorate UHI effects. As pilot studies, increased density of urban context and intensity of land use are deemed to influence on the livability of built-up environment through accelerating the formation of Urban Heat Island (UHI) effect (Stone and Rodgers 2001).

A great body of evidence has suggested the significant relationship between urban land use change and the dynamic of UHI. Methods and techniques have been largely developed respectively on the two data-exploration, i.e. land use change detection and UHI monitoring. However less has been done to detect the specific land surface changes, such as housing development, and thereafter correlate it with the pattern of UHI.

This paper would like to use satellite images and Hong Kong Statutory Planning information as data resource to extract the identifiable urban context of different years during 1987 to 2017 and explore the relationship between Hong Kong housing development and the dynamic of spatial pattern of UHI. The spatial patterns of UHI in the study years will be monitored based on the retrieval of Land Surface Temperature (LST) from the Landsat TM and OLI data. The research attempts to investigate how the spatial pattern of UHI relates to the changing of residential land use, and how it results from the internal housing renewal activities.

The research outcome of this paper primarily demonstrated on 2 phenomenons, i.e. dynamic of UHI in Hong Kong during 1987 and 2017, and the concurrent pattern of housing development, which can be the basis for peer researchers to explore the related topics. The correlative studies of housing development and UHI spatial pattern changes will be significant to advice the strategies to mitigate UHI effect through urban planning and design solutions, as well as provide a technical guidance to the future housing development of Hong Kong.

2 Methodology

2.1 Study Area

To specialize the interactions between housing developmental activities and dynamics of Urban Heat Island effects in Hong Kong, this research implemented the research methods on 4 targeted districts (Fig. 2), i.e. Yuen Long, Shatin, Kwun Tong, Kowloon City, which are the most representative new towns and old areas that undertakes tremendous housing developmental activities in Hong Kong.

Hong Kong has a hilly terrain situation. To cope with the mountainous environment and booming population, Hong Kong has promoted the development of new towns where appropriate in order to decentralize the population pressure of the old center. Since 1973, 9 new towns have been developed and gradually matured to bear the holistic function of a city and now are totally accommodating the majority of Hong Kong population. By 2016, the population of these nine new towns has approximated 3.47million, about 47.2% of total Hong Kong population at the time and is expected to rise to 3.63 million in 2021.

The Yuen Long district, the oldest district among Hong Kong's 18 districts, is composed of Yuen Long new town and Tin Sui Wai new town. Yuen Long district has a total population of 607200[1] by 2015.

The Shatin district is combined majorly by Shatin new town and Ma On Shan. The Shatin new town was built up majorly on reclaimed land and the development has been proceeded since the early 1970s. Till today, Shatin is accommodating about 659794 (see Footnote 1) people according to census.

The Kwun Tong district is one of the earliest developed areas and currently the most densely populated district of Hong Kong, with a density of 57530 people per square kilometer[2]. The total population of Kwun Tong is 648541 (see Footnote 1) by mid-2016.

The Kowloon City is another old district of Hong Kong. It locates in the center of Kowloon area. Residential is the dominant land use type and people mainly live in private housing. The total population of Kowloon City is 418723 (see Footnote 1) by mid-2016.

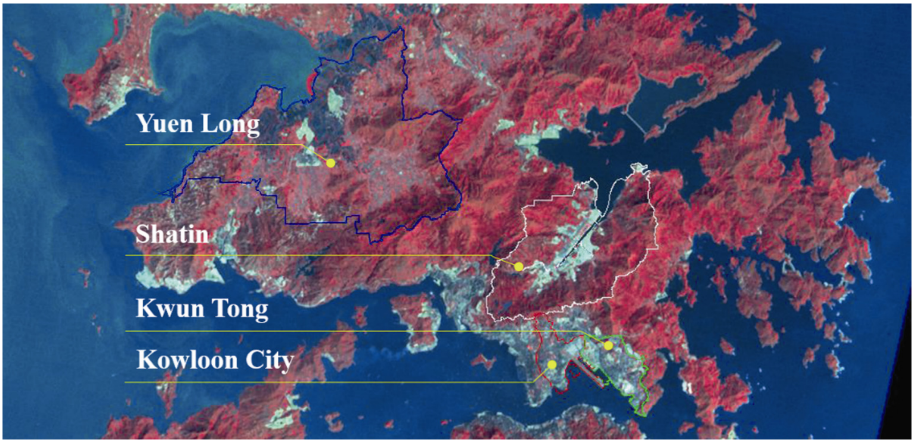


Fig. 2. Study areas - Yuen Long, Shatin, Kwun Tong and Kowloon City of Hong Kong

[1] Resource: *Hong Kong Population by Census 2016*. http://www.bycensus2016.gov.hk/tc/bc-dp.html

[2] Resource: *Hong Kong 2016: The Fact*. https://www.yearbook.gov.hk/2016/en/pdf/Facts.pdf

2.2 Data Resource

Remote sensing images were processed to detect the dynamic of urbanization and land surface temperature. To retrieve the changing of land surface information of the study duration, the remote sensing image of 4 years during 1987–2017 were utilized to imply the temporary transformation so that the dynamic could be more precisely interpreted. To eliminate the influence of season to temperature retrieval, remotely sensed imagery maps obtained in the adjacent months are adopted. The data utilized for this research are listed in Table 2.

Table 2. Remotely sensed imageries utilized in this research

Year	Acquisition date	Sensor type	Resolution
1987	18 December 1987	Landsat TM	30 * 30 and 120 * 20 for thermal infrared bands
1997	1 November 1997	Landsat TM	
2008	1 December 2008	Landsat TM	
2017	23 October 2017	Landsat OLI	30 * 30 and 100 * 100 for thermal infrared bands

The Landsat sensors record land surface information through the unique emitted energy of different land surface objects. The Landsat sensors have variety of bands which are distinguished by the special wavelengths of electromagnetic spectrum, e.g. Landsat TM has 7 bands and Landsat OLI has even 11 bands. The Landsat data represents land surface information in 30 * 30 pixels and each pixel contains a digital value within 0–255. The different land surface objects are identified through band math that highlights the objects based on calculated digital value range.

2.3 Data Processing

The accuracy of remotely sensed imagery to represent land surface information is usually influenced by atmosphere condition and sensor itself during the data acquisition progress. The Landsat data obtained in this research have been proceeded with radiometric calibration and to eliminate the influences of sensor errors and assure the accurate radiometric value acquired at the entrance of sensor. Afterward, the 4 groups of Landsat data were proceeded with atmospheric correction to eliminate the influence of errors caused by atmospheric scattering, absorption and reflection.

False Color Composites (FCCs) were produced to digitally enhance imagery for visual interpretation of spatial attributes and ground features (Lillesand et al. 2014). These FCCs are conducive to identify land cover classes such as urban, green space and water. Aided by the Geographical Information System technology, shapefile documents of the 4 study areas were created after spatial adjustment. Load and convert the shapefiles into research of interests (RoIs), and then overlap with the Landsat imagery. Finally subset the images with the respective RoIs to acquire the highlighted study areas.

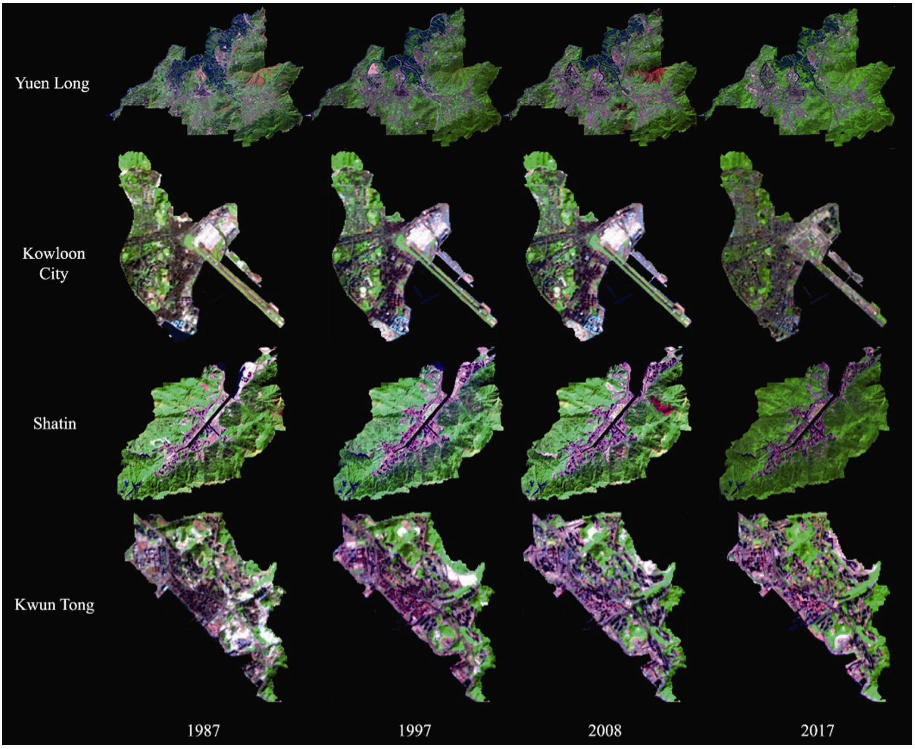


Fig. 3. False color composite images of Landsat TM and OLI NIR, green and blue bands to visually interpret urban transformation (Color figure online)

2.4 Calculation of Land Surface Temperature

The thermal infrared bands of Landsat TM, ETM+ and OLI were utilized to derive the land surface temperature. The methods to retrieve land surface temperature are differentiated by data resource, mainly including radiative transfer equation, split-window algorithm, temperature/emissivity separation method, mono-window method and single-channel method, etc.

As the Landsat TM sensor has only one thermal infrared band, the mono-window algorithm is applied to calculate brightness temperature, whilst radiation transfer equation is applied to Landsat OLI obtained data.

The general procedure to retrieve land surface temperature contains 4 steps: (1) transforming the digital value of thermal infrared band to thermal radiation brightness; (2) calculating land surface emissivity; (3) calculating the brightness temperature through thermal radiation intensity; and (4) calculate land surface temperature. The procedure can be shown as below (Figs. 4, 10 and 12).

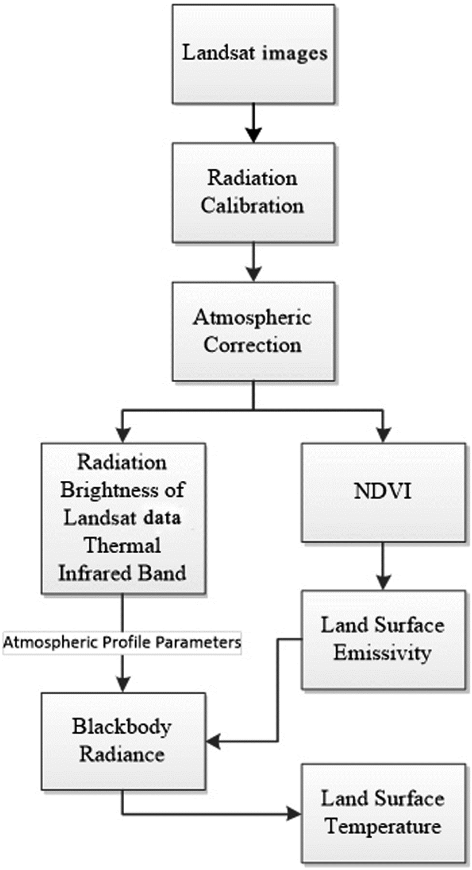


Fig. 4. General procedure of data processing

Firstly, the image should be pre-processed to get accurate radiative information. Radiative calibration is operated by applying calibration parameters. Equation (1) is applied to transforming the digital value of thermal infrared band to thermal radiation brightness:

$$L_\lambda = L_{\min(\lambda)} + \left(L_{\max(\lambda)} - L_{\min(\lambda)}\right) \times Q_{dn} \big/ Q_{max} \tag{1}$$

$$L_{\max(\lambda)} = 1.56\,\mathrm{m} \times \mathrm{W} \times \mathrm{cm}^{-2} \times \mathrm{sr}^{-1} \times \mu\mathrm{m}^{-1}$$

$$L_{\min(\lambda)} = 0.1238\,\mathrm{m} \times \mathrm{W} \times \mathrm{cm}^{-2} \times \mathrm{sr}^{-1} \times \mu\mathrm{m}^{-1}$$

$$Q_{mzx} = 255$$

Where L_λ refers to the atmospheric radiance received by the sensor; $L_{\max(\lambda)}$ is the maximum atmospheric radiance received by the sensor; $L_{\min(\lambda)}$ is the minimum atmospheric radiance received by the sensor; Q_{max} is the maximum digital value of pixels of the thermal infrared band; Q_{dn} is the digital value of pixels of the thermal infrared band.

To remove the scattering and absorption effects from the atmosphere to obtain the surface reflectance characterizing, an improved dark-object subtraction (DOS) is used during atmospheric correction. This method uses the shaded vegetation in the mountainous area as a black body, because the reflectance of the vegetation in the visible and mid-infrared band is very small, and the pixels in the shaded area have insufficient illumination, meantime it is difficult for the remote sensor to detect light in the shadow area (Liu et al. 2005).

The land surface emissivity is calculated based on Normalized Difference Vegetation Index (NDVI) which is obtained through Eq. (2). There's a significant relationship between NDVI value and the land surface emissivity value (Griend and Owe 1993; Qin et al. 2001). In this research, land surface emissivity is calculated following the NDVI range and corresponding emissivity values of Table 3 (Griend and Owe 1993; Qin et al. 2001; Liu and Zhang 2011).

$$\text{NDVI} = (\text{NIR} - \text{R})/(\text{NIR} + \text{R}) \tag{2}$$

Table 3. NDVI range and corresponding emissivity values

NDVI value	Surface emissivity
NDVI < −0.185	0.99
−0.185 ≤ NDVI < 0.157	0.970
0.157 ≤ NDVI ≤ 0.727	1.0094 + 0.047ln(NDVI)
NDVI > 0.727	0.990

The general concept of radiation transfer equation is to obtain the land surface radiance through eliminating the atmospheric influences from the at-sensor radiance and then transform the land surface radiance into the corresponding land surface temperature. Therefore, it's also necessary to obtain the blackbody radiance of thermal infrared band under the temperature T_s for the Landsat OLI data. The thermal infrared radiation brightness value received by the satellite sensor is composed of three parts, including the upward radiance from the atmosphere, the real radiance of the ground reaching the satellite sensor after passing through the atmosphere and the atmosphere radiates the energy reflected after reaching the ground. So the thermal infrared radiation brightness value received by the satellite can be written as the Eq. (3).

$$L_\lambda(T_s) = [\varepsilon B(T_s) + (1 - \varepsilon)L_\lambda \downarrow]\tau + L_\lambda \uparrow \tag{3}$$

Where, $L_\lambda(T_s)$ is the thermal infrared radiation brightness values received by satellite sensors at T, $L_\lambda \uparrow$ is the is the upwelling atmospheric radiance, $L_\lambda \downarrow$ is the downwelling atmospheric radiance, Then the blackbody radiance of thermal infrared band at T, which is $\mathrm{B}(T_s)$, can be calculated as follows.

$$B(T_s) = \frac{\mathrm{L}\lambda - \mathrm{L}\uparrow - \tau(1-\varepsilon)\mathrm{L}\downarrow}{\tau\varepsilon} \tag{4}$$

$$w = 0.0981 \times \left\{ 10 \times 0.6108 \times exp\left[\frac{17.27 \times (T_0 - 273.15)}{273.3 + (T_0 - 273.15)}\right] \times RH \right\} + 0.1697 \tag{5}$$

$$\tau = 1.031412 - 0.11536 \times \mathrm{w} \tag{6}$$

w is the water vapor content of atmosphere (g/cm2), τ is the atmospheric transmissivity, RH is the mean relative humidity (see Table 4), T_0 is the mean temperature of near surface level (see Table 4), ε is the land surface emissivity.

Table 4. Daily extract of meteorological observations of the data acquisition dates (Resource: Hong Kong Observatory Website. http://www.hko.gov.hk/cis/climat_e.htm)

	8 Dec 1987	1 Nov 1997	1 Dec 2008	23 Oct 2017
Absolute daily max T. (deg. C)	17.8	23	21.6	27.7
Absolute daily min T. (deg. C)	11.5	18.2	16	20.8
Mean T. (deg. C)	14.6	20.9	18.3	23.7
Mean relative humidity (%)	50	49	64	62

Brightness temperature is calculated based on thermal radiation intensity, following the Planck's function which is approximated to Eq. (6). The brightness temperature is the corresponding temperature of at-sensor radiance. This temperature doesn't present the real land surface temperature since the process is influenced by atmosphere condition and land surface roughness. However, the brightness temperature is a necessary variable to calculate the factual land surface temperature.

$$T_6 = K_2 / \ln\left(1 + K_1 / B_{(\lambda)}\right) \tag{7}$$

Where T_6 is the brightness temperature of digital pixels; K_1 and K_2 are constants.

For TM, $K_1 = 60.776\,\mathrm{m} \times \mathrm{W} \times \mathrm{cm}^{-2} \times \mathrm{sr}^{-1} \times \mu\mathrm{m}^{-1}, K_2 = 1260.56\,\mathrm{K}$.

For ETM+, $K_1 = 666.09\,\mathrm{m} \times \mathrm{W} \times \mathrm{cm}^{-2} \times \mathrm{sr}^{-1} \times \mu\mathrm{m}^{-1}, K_2 = 1282.71$ K.

For TIRS Band10, $K_1 = 774.89\,\mathrm{m} \times \mathrm{W} \times \mathrm{cm}^{-2} \times \mathrm{sr}^{-1} \times \mu\mathrm{m}^{-1}, K_2 = 1321.08$ K.

According to mono-window algorithm, Eq. (7) is applied to retrieve the land surface temperature based on remotely sensed imagery obtained by Landsat TM sensor, whilst the land surface temperature based on Landsat OLI is retrieved through

$$T_s = \{a(1 - C - D) + [b(1 - C - D) + C + D]T_6 - DT_6\} / C \tag{8}$$

$$T_a = 17.9769 + 0.91715 \times T_0 \tag{9}$$

$$a = -67.355351;\ b = 0.458606$$

$$C = \varepsilon \times \tau$$

$$D = (1 - \tau)[1 + (1 - \varepsilon) \times \tau]$$

Where, T_s is the land surface temperature (K); T_0 is the average mean atmosphere temperature (see Table 4).

2.5 Calculation of Urban Heat Island Intensity

To normalize the significance of land surface temperature value, the research use the surface temperature deviation from the mean value to estimate the Urban Heat Island Intensity.

$$\text{UHI}\ (\Delta T) = \frac{LST_s}{LST_a}$$

Where, UHI (ΔT) is the Urban Heat Island Intensity, LST_s is the retrieved land surface temperature and LST_a is the mean value of retrieved land surface temperature.

Assisted by Geographical Information Technology, the remotely sensed imagery was resampled to 30 m * 30 m vector grids to facilitate the calculation of Urban Heat Island Intensity changes through field calculation.

3 Result and Discussion

Comparing the dynamic of Urban Heat Island Intensity and the land surface cover change interpreted by the False Color Composites (see Fig. 3), it's manifested that the Urban Heat Island effects are intensified with the urbanization progress. All of the 4 study areas have different levels of intensification on the Urban Heat Island effects. Based on the unique urbanism activities happen in different regions, the Urban Heat Islands differ by years in morphology as amplified along with the sprawl of built-up areas or in the distribution as the changing of urban development focus.

3.1 Yuen Long District

Before 1987, when Yuen Long is in the beginning stage of development, the majority area of Yuen Long are covered by natural and agricultural landscapes. As shown in Fig. 5, there are two places with significant changes of Urban Heat Island distribution. The right area named "Pat Heung", which covering an area of 6 km^2, has long history and used to be the most populated region in Yuen Long. With the development of new towns, the Pat Heung attracted a large amount of immigrants, which resulted in the comparatively high living density and the higher intensity of Urban Heat Island as detected. With the social-economic development, Yuen Long also developed local industry following the new town policies. The left circle shows the industry area of Yuen Long, which covers area of 67 hector and established since 1983. The industry emission and human activities within the industry area raise up the intensity of Urban Heat Island effect.

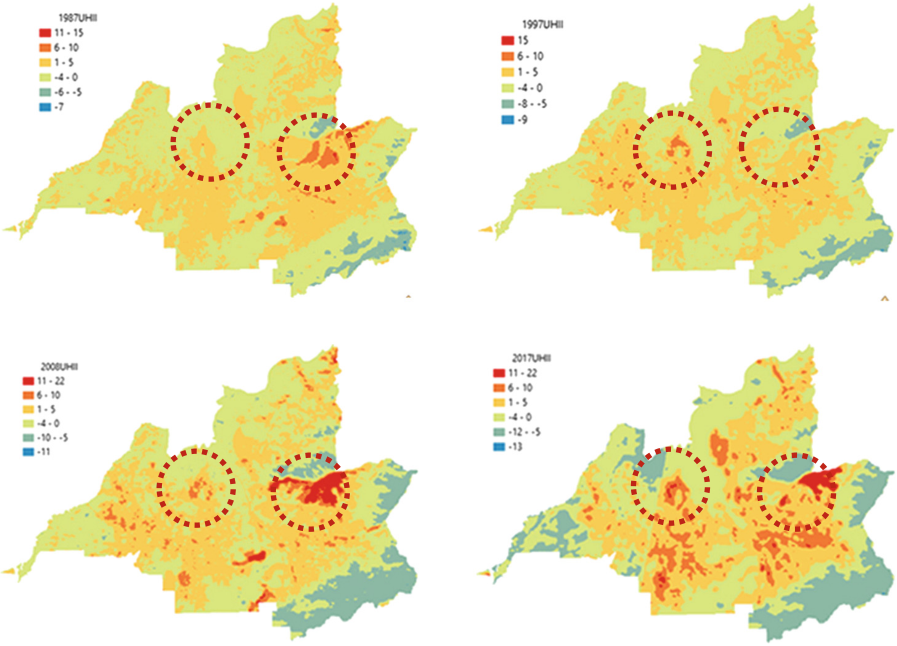


Fig. 5. Urban heat island intensity of Yuen Long District from 1987 to 2017

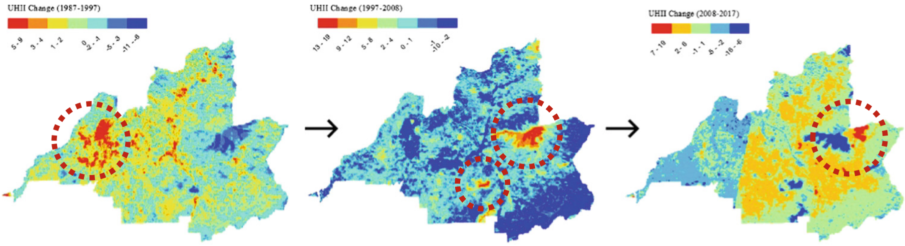


Fig. 6. Dynamic of UHII of Yuen Long District of 1987–1997, 1997–2008 and 2008–2017

Comparing the images of 1997 and 1987, there is a significant intensification of Urban Heat Island effect in 1997 as the red and yellow areas are larger than before. It related to the development of Tin Sui Wai new town. Before the 1980s, The Tin Sui Wai area is covered by mangrove forest, which was filled during 1987 to 2000. Since 1990, the Hong Kong government started to develop the southern area that the first public housing, established in 1992, and private housing projects, established in 1991, all locate in the southern area.

The circle area in the Fig. 7 indicates the major road system of Yuen Long new town and the surrounding areas, including Yuen Long High Way and Castle Peak Road which were put into use respectively in 1994 and 1998. The enriched road system facilitates the development of new town and thus the Urban Heat Island effect was significantly increased. Besides, there are a number of housing projects and commercial projects rising up in the nearby areas, including the Yuen Long Center built up in 1993, the Greenfields built up in 1999 and the Grand Del Sol built up in 1997 and the Yuen Long Plaza built up in 1989, etc. In another word, Yuen Long new town encountered the peak of development during 1987 to 1997 and the population has also increased form the 211,540 in 1986 to 341,030 in 1996.

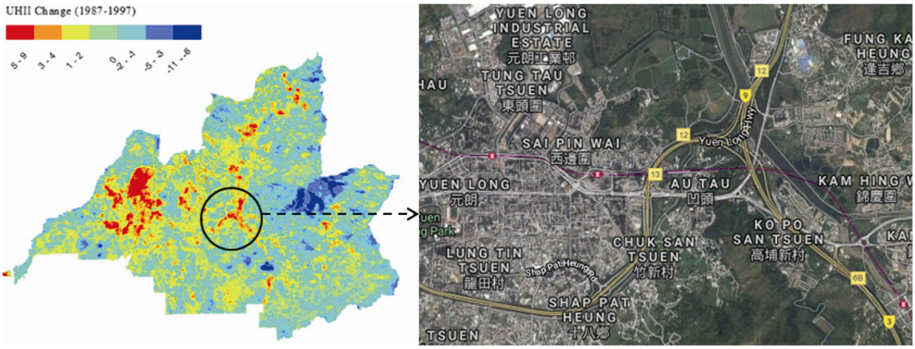


Fig. 7. Transportation system related UHI in Yuen Long

During 1997 to 2008, excluding some of the areas were intensified of Urban Heat Island effect (red areas as shown in Fig. 6), the Urban Heat Island condition of other areas maintain the same situation or was mitigated to some extends. During this period, urban greening was paid increasing attention in the Tin Sui Wai and Yuen Long new towns, which is a unique and leading action among all the new towns of Hong Kong. This is proposed as one of the reason of the UHI mitigation when the population of Yuen Long kept increasing form 341,030 of 1996 to the 534,192 in 2016.

During 2008 to 2017, the Urban Heat Island effect in Yuen Long maintains stable. The two new town in Yuen Long District, Yuen Long and Tin Sui Wai, served as engines to promote the development of surrounding areas in the past 3 dates, such as the large yellow and orange areas. Since housing provision is still challenging Hong Kong, the Hong Kong government made much efforts to improve land use efficiency to complement sufficient housing units and infrastructure while kept the conservation of country parks and farmland.

3.2 Shatin District

In the early 1970s, Shatin is a town with population of 30000. Since the establishment of Shatin new town in 1995, Shatin has become the second largest new town in Hong Kong (Deng 1995). Before 1987, a great number of building and housing projects were built in Shatin, such as the Shatin New Town Plaza in 1984, the Lek Yuen Estate in 1977, the Shatin Center in 1981, the Lucky Plaza in 1984, and the City One during 1982 to 1988, etc. The both sides of Shing Mun River are reclaimed land, which have been mainly for residential development use. A great number of early housing projects of

Shatin new town located along the Shing Mun River. This is proposed as one of the reasons that the both sides of the Shing Mun River appear to have higher Urban Heat Island Intensity (See Fig. 8).

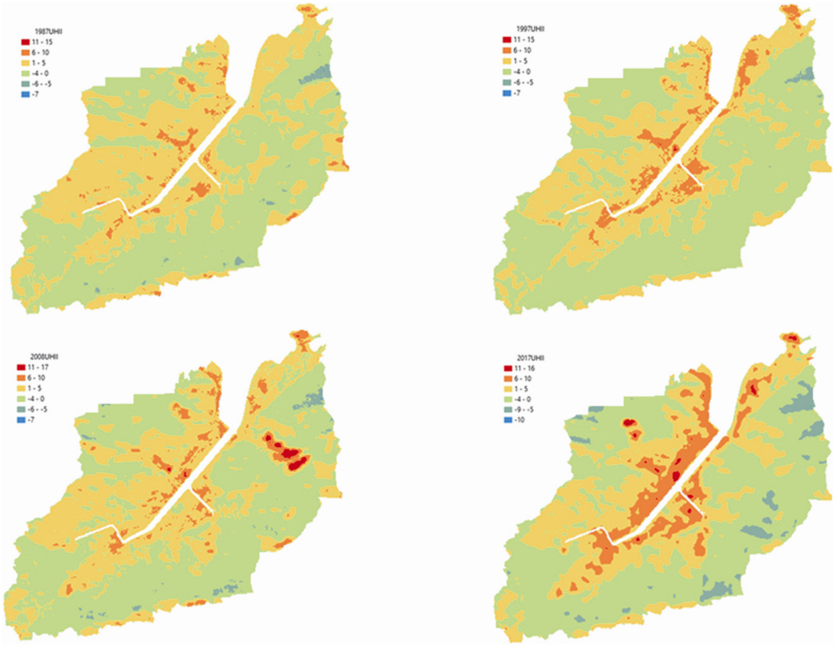


Fig. 8. Dynamic of urban heat island intensity of Shatin District from 1987 to 2017

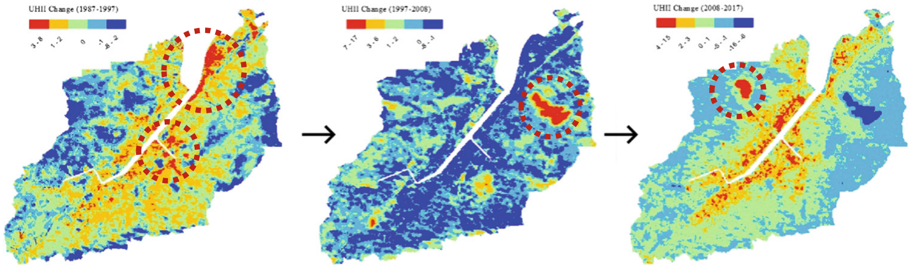


Fig. 9. Dynamic of UHII of Shatin District of 1987–1997, 1997–2008 and 2008–2017

During 1987 to 1997, the Urban Heat Island distributed in Shatin are intensified, according to the expanded orange and red areas in the Fig. 8. The circled areas are the most significant ones according to Fig. 9. According to statistics, the total population of Shatin increased 61% with immigrants form New Territories, Kowloon Island and Hong Kong Island. To satisfy the housing demands of increasing population, tremendous housing projects were planned and implemented during the period, which may become a key reason to intensify the regional Urban Heat Island effect. Comparing the changes of Urban Heat Island Intensity during 2008 to 2017, there's no significant intensification of UHI with the cooling-down of urbanization progress.

3.3 Kwun Tong

Compare with Yuen Long and Shatin, the distribution of Urban Heat Island is not that significantly changed. Referencing to the implementations regarding to urban planning and design, there are several events considered to contribute to the changes of Urban Heat Island in the Kwun Tong area.

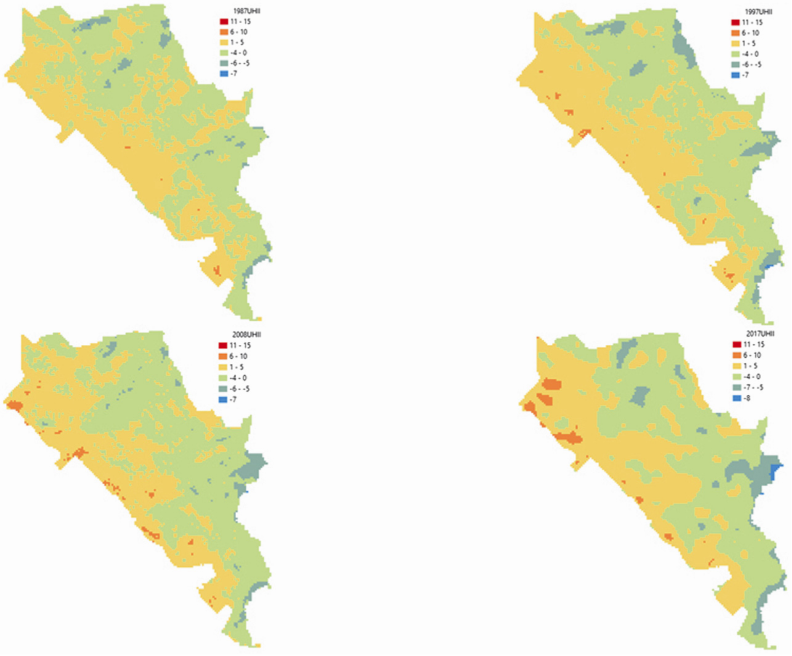


Fig. 10. Urban heat island intensity of Kwun Tong District from 1987 to 2017

During 1998 to 1999, the Hong Kong Government redesigned the staff facilities, facades, ticketing hall and platform of Kwun Tong Station to comply with the Airport Express Link, installing air-conditioning system, changing the wall and column of the old dark gray pebble stone into white man-made fiber board, and also adding lifts to and from the ground. This leads to a tentative high land surface temperature as detected. The renovation of pedestrian bridge superstructure started in 2008. The new cover of the material is divided into two types: metal plate and transparent fiberboard. Metal plate was placed in the middle of the cover to protect the lights, speakers and other power facilities, transparent fiber board was placed on both sides of the cover to increase daylighting and, reduce the use of electric light. The environmental-friendly renovation implementations is conducive to mitigate the regional Urban Heat Island effect.

In 1978, the Hong Kong government established a modern landfill at Sai Tso Wan called Shatuwan Garbage Landfill with an area of about 9 hectares and piled over 1.6 million tons of waste. The deepest thickness of 65 m. The landfill was closed in 1980 and then left vacant for more than 10 years. Until 1997, the landfill was undergone restoration work. The 2.8 hectares of land was rebuilt into the Sai Tso Wan Recreation Ground in 2004. It was the first permanent recreation facility in Hong Kong built on a restored landfill. As the green circles in Fig. 11, the progress is proposed to be one of the reasons to increase the Urban Heat Island intensity, since the un-used bare land may also contribute to the formation of UHI.

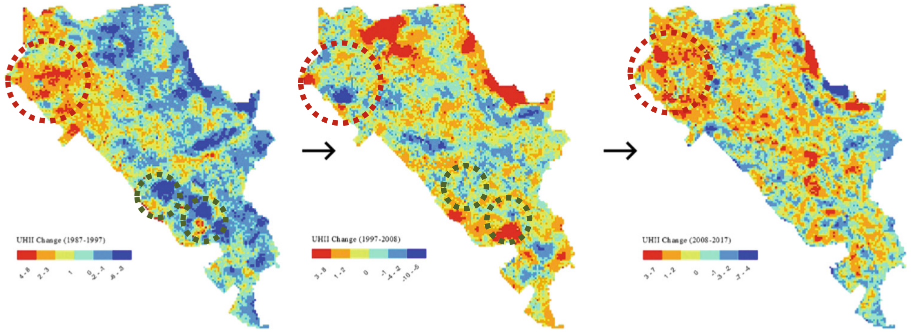


Fig. 11. Dynamic of UHII of Kwun Tong District of 1987–1997, 1997–2008 and 2008–2017

3.4 Kowloon City

According to the Urban Heat Island detection result, Kai Tak area has higher temperature than the other places of Kowloon City. In July 1998, the Hong Kong International Airport was relocated from the densely populated area of Kowloon City to Chek Lap Kok on Lantau Island. The Kai Tak Development Project (see red circles in Fig. 13), a large-scale urban development project conducted by the Hong Kong Government at the site of Kai Tak Airport, was launched. In 2006, the Planning Department (UNDP) implemented the Development Outline and planned to construct two public housing estates in the Kai Tak Development Area. The overall design of the PRH project was planned by the Chief Architect of the Housing Department, SAR. Affiliated company construction. In 2008, in response to the objection raised against the flat screen at the time, the Housing Department widened the building to enhance ventilation and greenery. As a result, about 30% of the dug pile had been laid off to rebuild the foundation at a cost of about $ 400 million. To 570,000 square feet, floor height reduced to 35–40.

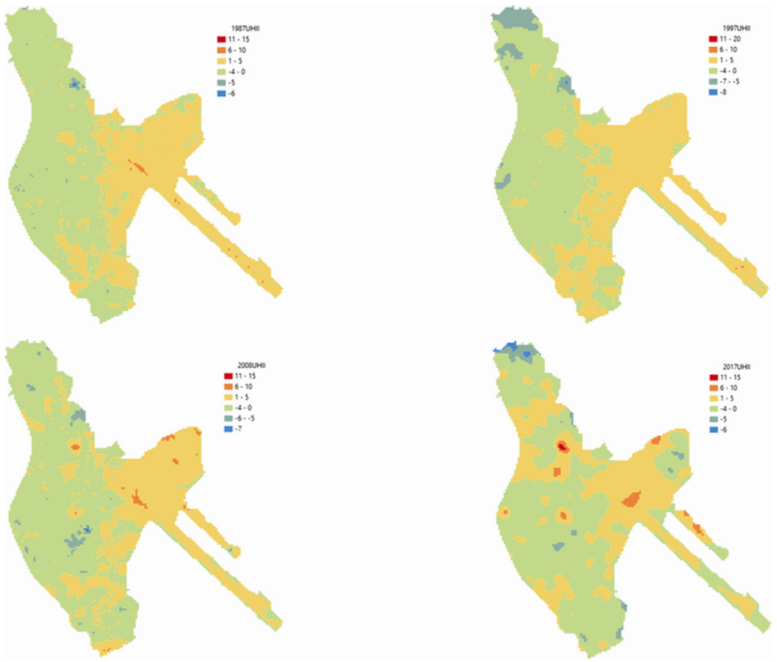


Fig. 12. Urban heat island intensity of Kowloon City District from 1987 to 2017

In the mid-1990s, the station platform was extended to Hung Hom Bay (see blue circles in Fig. 13). The original bus terminal on the platform was relocated to the present site. The new platform also replaced the old lobby with a new lobby designed by British architect Lord Norman Foster and completed in 1997. The new lobby has a wavy roof design with turquoise steel to the south and glass skylights to the north and 3 glass facades.

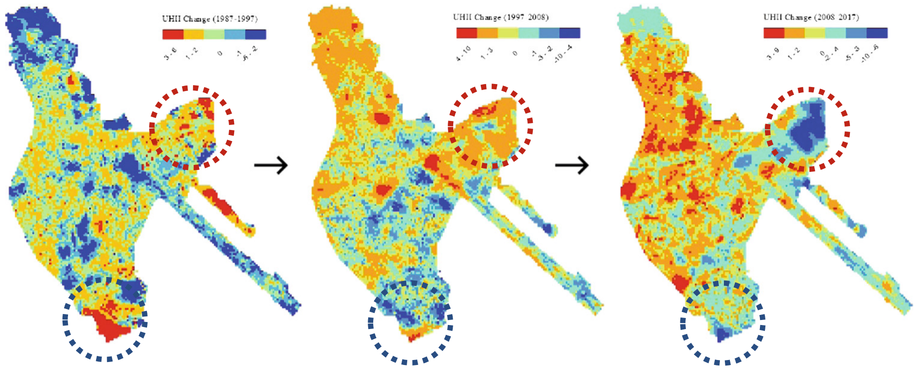


Fig. 13. Dynamic of UHII of Kowloon City District of 1987–1997, 1997–2008 and 2008–2017 (Color figure online)

4 Conclusions

In this study, the mono-window algorithm and radiation transfer equation are utilized to retrieve land surface temperature of 4 areas of Hong Kong, i.e. Yuen Long, Shatin, Kwun Tong and Kowloon City, in the year 1987, 1997, 2008 and 2017, using remotely sensed imagery obtained by Landsat TM and Landsat OLI. Based on the acquired land surface temperature, Urban Heat Island Intensity is further calculated to interpret the intensification or mitigation of regional Urban Heat Island effect.

Through the retrieved land surface temperature and calculated Urban Heat Island Intensity, it's found all of the four regions encountered the intensification of Urban Heat Island in the progress of urbanization. Housing development and related construction activities contribute to the dynamic of Urban Heat Island effect. The un-used bare land which is left exposed to solar radiation and the utilization of building materials with high albedo rate may intensify the regional Urban Heat Island effect. Maintaining natural resources, like green space and urban farmland will be helpful to mitigation. In this regard, it's suggested to maintain reasonable construction progress to avoid expose bare land for long time to the solar radiation. In addition, it's also important to plan and design urban green space based on the ecological and environmental function, excluding the landscape function.

This research still has some limitations. Since housing development is a complex process, the social and economic factors should also be taken into consideration when evaluating the environmental impact of housing. Moreover, the vertical dynamics of housing is not discussed in this research but should be considered and quantified on the impacts to Urban Heat Island in the further research.

Acknowledgement. This research was supported by the Faculty of Social Science Direct Grant on the project titled "Study of impact of hyper-density housing development on Urban Heat Island of Hong Kong".

References

Chun, B., Guldmann, J.-M.: Spatial statistical analysis and simulation of the urban heat island in high-density central cities. Landsc. Urban Plan. **125**, 76–88 (2014)

Deng, W.: The new town development and planning in Hong Kong. Urban Plann. Int. (4), 7–11 (1995)

Giridharan, R., Lau, S.S.Y., Ganesan, S., Givoni, B.: Urban design factors influencing heat island intensity in high-rise high-density environments of Hong Kong. Build. Environ. **42**(10), 3669–3684 (2007). https://doi.org/10.1016/j.buildenv.2006.09.011

Griend, A.A.V.D., Owe, M.: On the relationship between thermal emissivity and the normalized difference vegetation index for natural surfaces. Int. J. Remote Sens. **14**(6), 1119–1131 (1993). https://doi.org/10.1080/01431169308904400

Hiemstra, J.A., Saaroni, H., Amorim, J.H.: The urban heat island: thermal comfort and the role of urban greening. In: Pearlmutter, D., Calfapietra, C., Samson, R., O'Brien, L., Krajter Ostoić, S., Sanesi, G., Alonso del Amo, R. (eds.) The Urban Forest. FC, vol. 7, pp. 7–19. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-50280-9_2

Kyriakodis, G.E., Santamouris, M.: Using reflective pavements to mitigate urban heat island in warm climates-results from a large scale urban mitigation project. Urban Clim. (2017)

Li, J., Wang, X., Wang, X., Ma, W., Zhang, H.: Remote sensing evaluation of urban heat island and its spatial pattern of the Shanghai metropolitan area. China Ecol. Complex. **6**(4), 413–420 (2009)

Lillesand, T., Kiefer, R.W., Chipman, J.: Remote Sensing and Image Interpretation. Wiley, Chichester (2014)

Liu, L., Zhang, Y.: Urban heat island analysis using the Landsat TM data and ASTER data: a case study in Hong Kong. Remote Sens. **3**(7), 1535–1552 (2011)

Ng, E., Ren, C.: The Urban Climatic Map: A Methodology for Sustainable Urban Planning. Routledge, London (2015)

Qin, Z., Karnieli, A., Berliner, P.: A mono-window algorithm for retrieving land surface temperature from Landsat TM data and its application to the Israel-Egypt border region. Int. J. Remote Sens. **22**(18), 3719–3746 (2001)

Santamouris, M.: Energy and Climate in the Urban Built Environment. Routledge, New York (2013)

Stone Jr., B., Rodgers, M.O.: Urban form and thermal efficiency: how the design of cities influences the urban heat island effect. J. Am. Plan. Assoc. **67**(2), 186–198 (2001)

Streutker, D.R.: A remote sensing study of the urban heat island of Houston, Texas. Int. J. Remote Sens. **23**(13), 2595–2608 (2002). https://doi.org/10.1080/01431160110115023

Weng, Q., Lu, D., Schubring, J.: Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies. Remote Sens. Environ. **89**(4), 467–483 (2004). https://doi.org/10.1016/j.rse.2003.11.00

Wong, M.S., Peng, F., Zou, B., Shi, W.Z., Wilson, G.J.: Spatially analyzing the inequity of the Hong Kong urban heat island by socio-demographic characteristics. Int. J. Environ. Res. Public Health **13**(3), 317 (2016)

Yuan, F., Bauer, M.E.: Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. Remote Sens. Environ. **106**(3), 375–386 (2007). https://doi.org/10.1016/j.rse.2006.09.003

Zhang, J., Wang, Y., Li, Y.: A C++ program for retrieving land surface temperature from the data of Landsat TM/ETM+ band6. Comput. Geosci. **32**(10), 1796–1805 (2006). https://doi.org/10.1016/j.cageo.2006.05.001

Zhang, Y.: Detection of urban housing development by fusing multisensor satellite data and performing spatial feature post-classification. Int. J. Remote Sens. **22**(17), 3339–3355 (2001)

Liu, X.P., Deng, R.R., Peng, X.J.: A fast atmospheric correction method based on TM imagery. Sci. Geogr. Sinica **25**(1), 87–93 (2005)

PPP Cost-Sharing of Multi-purpose Utility Tunnels

Ali Alaghbandrad[1] and Amin Hammad[2(✉)]

[1] Department of Building, Civil and Environmental Engineering, Concordia University,
1515 Sainte-Catherine Street West, Montreal, QC H3G 2W1, Canada
alag_ali@encs.concordia.ca
[2] Concordia Institute for Information Systems Engineering, Concordia University,
1515 Sainte-Catherine Street West, Montreal, QC H3G 2W1, Canada
hammad@ciise.concordia.ca

Abstract. Construction, maintenance, and renewal of underground utilities (e.g. gas, water and sewer pipes, and electrical and telecommunication cables) impose a large lifecycle cost on utility companies and the citizens of urban areas. Integrating all the utilities in a Multi-purpose Utility Tunnel (MUT) can reduce post-construction costs of accessibility, inspection, maintenance, protection, and social costs. Although the high construction cost is an obstacle for promoting MUTs, post-construction cost savings make MUTs an economic alternative solution considering the total lifecycle costs. On the other hand, sharing the lifecycle cost of MUTs is another issue, particularly for the high initial investment in construction. Public-Private Partnership (PPP) is a promising approach for sharing the lifecycle cost of MUTs among public and private investors and facilitating the development of MUT projects.

This paper aims to develop a model for PPP financing and cost-sharing of MUT projects using game theory. The PPP investors need to investigate different scenarios for lifecycle cost-sharing and choosing the best strategy considering their financial situation. The proposed model is based on three steps: (a) selection of MUT cost allocation method for MUT lifecycle phases, (b) cost adjustment based on risk and benefit cost factors, (c) game model of MUT cost-sharing based on an entity covering a part of the MUT construction cost using loans/bonds with a penalty that is used as a reward to the other entity. The results show that this model can improve the fairness of MUT cost allocation and be used as an analytical tool for the MUT project stakeholders to choose appropriate financing strategy based on anticipated lifecycle benefits and costs.

Keywords: Multi-purpose Utility Tunnel · Public-Private Partnership
Game theory

1 Introduction

The traditional method of buried utilities is common for the development of utility networks, such as gas, water and sewer pipes, and electrical and telecommunication cables, especially in urban areas. Although relatively low initial construction cost of

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 554–567, 2018.
https://doi.org/10.1007/978-3-319-91635-4_29

buried utilities is the main advantage of this method, the lifecycle cost analysis shows it is equal or even a more expensive method [1, 2]. The major post-construction cost is related to accessibility to the utilities for maintenance activities (i.e., inspection, repair, and renewal). The repeated excavation of roads to access the utilities during the operation phase imposes direct construction cost on the utility companies and social costs on the citizens. The social costs include a wide variety of costs, such as [2] (a) vehicle and pedestrian delay cost due to traffic congestion, (b) health costs due to environmental pollution, (c) decreased productivity costs as a result of dust, noise, and vibration, (d) local business loss as a consequence of excavation, and (e) safety costs.

Another indirect post-construction cost for utility companies is related to increased repair and decreased lifespan of buried utilities, as a result of unprotected underground space. For example, soil humidity accelerates the corrosion of pipes, and water penetration disturbs the functionality of electricity and telecommunication cables. Accidental damage to underground networks in construction work is another costly issue of the unprotected underground space [2].

Even without considering the social costs, which are not easy to quantify, the lifecycle cost of buried utilities is high. An alternative solution for the buried utilities method is accommodating all the utilities in an isolated and protected underground space, i.e. Multi-purpose Utility Tunnels (MUTs).

However, the financing of MUT projects is a big issue [1, 2]. Although the high construction cost is an obstacle for promoting MUTs, post-construction cost savings make MUTs an economic alternative solution considering the total lifecycle costs. On the other hand, sharing the lifecycle cost of MUTs is another issue, particularly for the high initial investment in construction. Public-Private Partnership (PPP) is a promising approach for sharing the lifecycle cost of MUTs among public and private investors and facilitating the development of MUT projects.

This paper aims to develop a model for PPP financing and cost-sharing of MUT projects based on lifecycle cost analysis and different scenarios of cooperative/non-cooperative game theory. In addition, different methods of cost-sharing are discussed and cost adjustment methods are proposed to improve the fairness of cost allocation in this paper.

2 Literature Review

2.1 Multi-purpose Utility Tunnels

A Multi-purpose Utility Tunnel (MUT) is defined as "an underground utilidor containing one or more utility systems, permitting the installation, maintenance, and removal of the system without making street cuts or excavations" [3]. Using MUT for integrating municipal utility networks, such as electricity, telecommunication, water, sewerage, and gas, is beneficial for two groups: (a) utility providers, (b) utility users and citizens.

The main benefits for utility providers are: (a) elimination of costs related to repeated construction (excavation [3–5], utility placement [6], road and sidewalk repair [2], traffic control [7], detour road damage due to extra traffic load [8]), (b) improved inspection and maintenance of utilities [2, 6, 9], (c) minimization of damage [6] and corrosion of utilities [10], (d) future development and upgrade cost savings [9, 11], (e) elimination of labor accidental injury and death [9, 12], (f) elimination of municipal revenue loss [12, 13], and (g) more organized planning of underground space [14].

The main benefits of MUT for utility users and citizens (social benefits) are: (1) elimination of traffic congestion [7, 9, 12], (2) improved health, environment, and safety [7, 8, 12, 15], (3) improved quality of utility services and customer satisfaction [3, 4, 16], (4) elimination of local business loss [12, 17], and (5) elimination of damage/temporary closure of recreational facilities, e.g. parks [12].

The main disadvantages of MUTs are (1) high initial investment cost [2, 5], (2) compatibility and safety issues [3, 18], (3) security issues [6], (4) coordination issues [16], (5) disruption of services during the construction phase [3, 18], and (6) less-known construction methods [5].

MUTs are classified by type, installation place, shape, and material [5]. MUT types include searchable, visitable, compartmentalized; and can be installed under road, sidewalk, and metro. Different shapes of MUTs are trapezoid, rectangular, rectangular with lid, circular, ovoid with gutter, double oval. The tunnel material can be High-Density Polyethylene (HDPE), cast-in-place concrete, pre-cast concrete sections, steel, brick and mortar, sprayed concrete.

2.2 Economy, Financing, and Cost-Sharing of MUTs

MUT Economy. The initial construction cost of MUTs is higher than traditional buried utilities method. However, utility companies will have cost savings from MUT benefits mentioned in Sect. 2.1. These benefits will be obtained during the operation phase, and make the payback period of MUT very long. From a lifecycle perspective, there is a point that the total construction and operation cost of MUTs is equal to the traditional buried utilities method (open-cut) as shown in Fig. 1 [1].

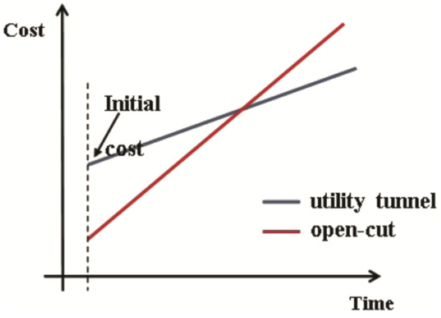


Fig. 1. Cost curve of open-cut and MUT [1]

After this time, MUT cost saving makes it more economical. In addition, more cost savings are obtained by adding social benefits [1]. There are other factors used for the economic evaluation of MUT compared to the traditional buried utilities method, such as the utility type, number of pipes (i.e. density), pipe diameter, number of excavation and reinstatement (E&R) avoided, location (i.e. undeveloped, suburban and urban areas), and the depth of the MUT [2].

Public-Private-Partnership (PPP) for MUT. The long-term economic perspective for MUT and cooperation between several public and private MUT stakeholders make PPP a practical form of contract for these projects [1]. PPP is defined as a cooperation between the public and private sectors in order to execute projects or deliver public services, which is traditionally provided by the public sector. Both the public and private sectors gain some benefits in proportion to the degree of their involvement in specific tasks. The main purpose of PPP is to assign the risk to the sector that can better control it [19]. China is one of the leading countries in building MUTs with a total length of MUTs built in major cities from 1994 to 2015 of about 500 km (Fig. 2).

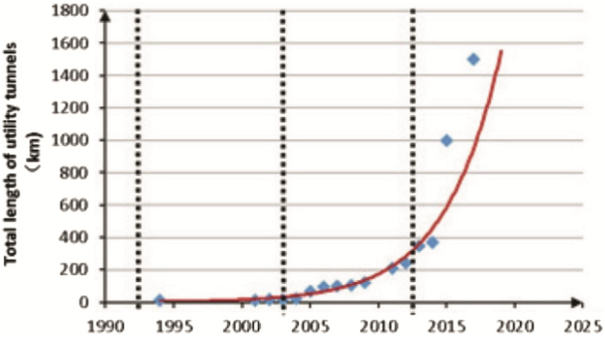


Fig. 2. Growth curve of MUT construction in China [1]

The Chinese government adopted the PPP approach to attract private capital to MUT construction projects. Using this approach enabled a long-term cooperation (e.g. 20 years) between the public and private investors and also sharing the risks and benefits of MUT projects. Figure 3 demonstrates the PPP model for MUT projects in China. In this model, the Chinese government only pays a low percentage of the construction cost of the MUT project. However, the government is responsible for providing a stable environment for the investment return of the private investors. For this purpose, the main tasks of the government are providing institution environment and legislative guarantee, such as the establishment of subsidy, supervision, and payment mechanisms [1].

The PPP model of MUT for each country can be adjusted by considering the specific conditions of the context, such as the government power for interventions and guarantees in MUT projects, the anticipated profit and risks of the MUT project, etc.

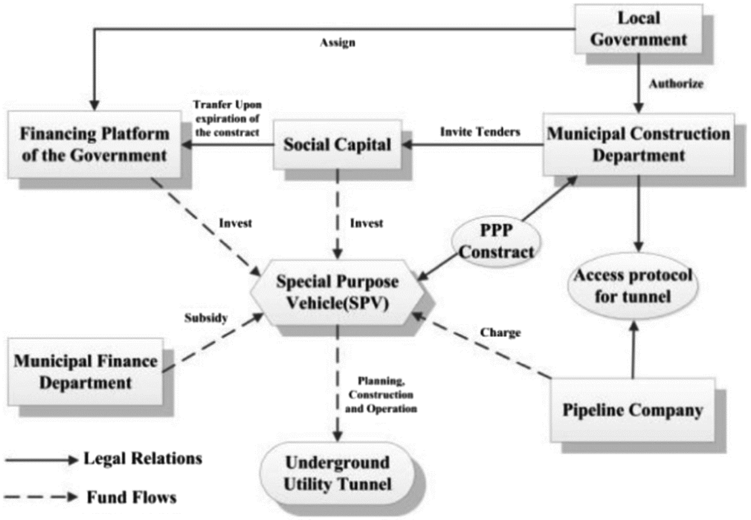

Local Government
Assign
Authorize
Tranfer Upon expiration of the constract
Financing Platform of the Government
Social Capital
Invite Tenders
Municipal Construction Department
Invest
Invest
PPP Constract
Special Purpose Vehicle(SPV)
Access protocol for tunnel
Subsidy
Charge
Municipal Finance Department
Planning, Construction and Operation
Pipeline Company
Legal Relations
Fund Flows
Underground Utility Tunnel

Fig. 3. PPP model for MUT projects in China [1]

Cost Allocation of MUT Projects. Utility companies, as the users of MUTs, are responsible for contributing the cost of these projects. Fairness of cost allocation for MUT projects is a challenging issue. Currently, the common two methods of MUT cost allocation are: (a) The proportion of buried cost (PBC) method, in which the utility companies are charged based on the same proportion they were paying in traditional buried utilities method [20, 21], and (b) The proportion of utility volume occupancy (PUVO) method, in which utility companies are charged based on the volume of space they occupy in MUT. A combination of these two methods is also proposed [21]. More details are provided in Sect. 3.1.

Game Theory for MUT Cost-Sharing. Game theory, also called "conflict analysis" or "interactive decision theory," is "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers" [22]. A method based on game theory for the cost-sharing strategy of MUT is proposed in China [23]. Game theory is applied to design a government incentive mechanism for financing MUT construction. The game is based on two utility companies making four possible scenarios of sharing or not sharing the construction cost. If both utility companies accept to share the construction cost, the cost is shared between them by a certain ratio. If one of the two companies does not agree to share the construction cost, but the other company agrees, a percentage of the share is reduced to the paying company as a reward and added to the other company as a penalty. In case both companies do not agree to share the construction cost, MUT still will be built anyways and both companies must pay a certain fee every year to place utility inside MUT [21].

Both of these game models for MUT cost-sharing are based on high degree of public entities intervention, such as government or municipality. Although the role of public entities in legislative and coordination affairs is undeniable, various financing options should be given to the other MUT stakeholders.

3 Proposed Method

The proposed method is based on three steps as shown in Fig. 4: (a) selection of MUT cost allocation method for MUT phases, (b) weighted factors adjustment, and (c) game theory.

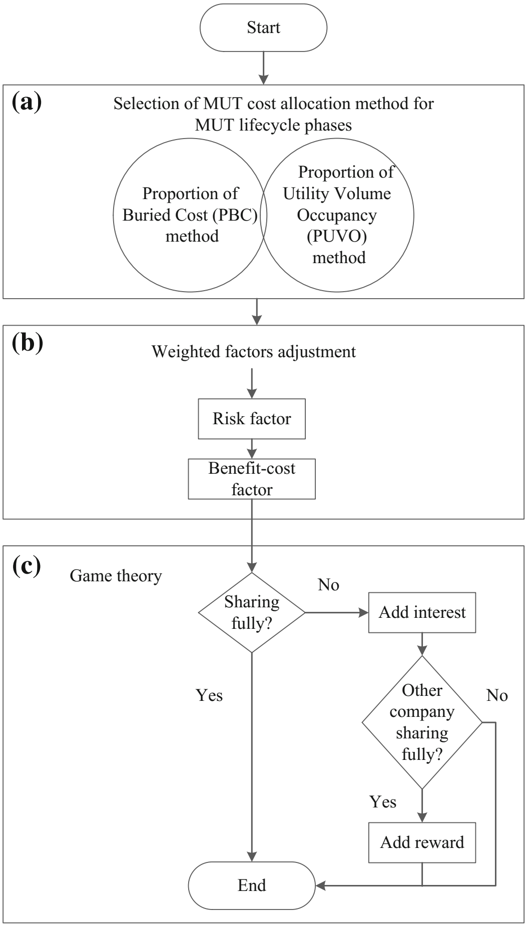

Start
(a)
Selection of MUT cost allocation method for MUT lifecycle phases
Proportion of Buried Cost (PBC) method
Proportion of Utility Volume Occupancy (PUVO) method
(b)
Weighted factors adjustment
Risk factor
Benefit-cost factor
(c)
Game theory
Sharing fully?
No
Add interest
Yes
Other company sharing fully?
No
Yes
Add reward
End

Fig. 4. Research method

3.1 Selection of MUT Cost Allocation Method

As explained in Sect. 2.2, two methods of MUT cost allocation have been proposed: the PBC method and the PUVO method. This paper proposes using the PBC method for the design and construction phases and the PUVO method for the operation phase of MUT. In the PBC method, the design/construction cost portion of the MUT for utility company $i\left(CM_i\right)$ can be calculated by Eq. 1.

$$CM_i = \frac{CB_i}{\sum_{i=1}^{n} CB_i} \times CM_t \tag{1}$$

where CM_t is the total design/construction cost, and CB_i is the design/construction cost of utility company i in the buried method.

In the PUVO method, the total lifecycle operation cost portion of the MUT for utility company $i\left(OCM_i\right)$ can be calculated by Eq. 2.

$$OCM_i = \frac{OV_i}{\sum_{i=1}^{n} OV_i} \times OCM_t \tag{2}$$

where OCM_t is the total operation cost and OV_i is the occupied volume of utility company i in the MUT.

The values of CM_i and OCM_i include only the common costs of MUT, e.g. tunnel and service equipment of MUT. The specific cost of each utility company, e.g. costs of material, installation of the network inside the MUT, maintenance of network, etc., are excluded from CM_i and OCM_i and will be paid directly by each utility company.

3.2 Weighted Factors Adjustment

Although the proportion of costs in the PBC and PUVO methods are the main factors for allocation of MUT cost to utility companies, there are other factors to be considered for cost allocation.

Risk Factor. Accommodation of certain utilities together imposes safety risks on the MUT. For example, the proximity of gas pipes and electrical wires in the MUT increases the risk of fire. In addition to the insurance cost for safety issues, an extra cost is needed for the safer design of the MUT and the installation and operation of safety devices, such as fire detectors and sprinklers. A fair cost-sharing method should allocate the costs of risks to the utility companies that bring the risks.

For risk adjustment at any phase of the MUT lifecycle, the total cost of risk k at phase p (CR_{kp}) should be deducted from the total costs (CM_t and OCM_t) and redistributed among utility companies based on the risk indicator of company i (γ_i) such that $\sum_{i=1}^{n} \gamma_i = 1$. The cost of risk k (CR_{kip}) for utility company i in phase p of the project can be calculated by Eq. 3:

$$CR_{kip} = \gamma_i CR_{kp} \tag{3}$$

Benefit-Cost Factor. As explained in Sect. 2, the benefits of MUT are considered during the operation phase. However, utility companies may not benefit equally. For example, a company may benefit largely from facilitated inspection, while other companies may benefit mainly from the elimination of repeated excavation or minimization of corrosion of utilities. Therefore, the benefit factor must be defined and allocated to each utility company by considering their conditions. The companies with larger gained benefit must contribute more to MUT costs.

For benefit adjustment, the MUT benefit-cost ratio for utility company *i* (BCR_i) should be calculated by dividing the total lifecycle benefits (MB_i) by the total lifecycle costs of MUT after risk adjustments using Eq. 4.

$$BCR_i = \frac{MB_i}{CM_i + OCM_i} \tag{4}$$

For the sake of fairness, the variance of BCR_i from the average MUT benefit-cost ratio (BCR_{ave}) should be less or equal to a threshold *T* (Eq. 5), which is already agreed on.

$$|BCR_i - BCR_{ave}| \le T \tag{5}$$

If this condition is not satisfied, there are two cases: (a) for companies that satisfy the condition $BCR_i < BCR_{ave} - T$, the company with the lowest *BCR* is selected, and a portion of the cost of this company (CT_i) is transferred to the company with the highest *BCR* to satisfy Eq. 5. This amount can be calculated by Eq. 6. This process is applied iteratively for all other companies satisfying the above condition.

$$CT_i = CM_i + OCM_i - \frac{MB_i}{BCR_{ave} - T} \tag{6}$$

(b) for companies that satisfy the condition $BCR_i > BCR_{ave} + T$, the company with the lowest *BCR* is selected, and a portion of the cost of this company (CT_i) is transferred to the company with the highest *BCR* to satisfy Eq. 5. This amount can be calculated by Eq. 7. This process is applied iteratively for all other companies satisfying the above condition.

$$CT_i = \frac{MB_i}{BCR_{ave} + T} - CM_i - OCM_i \tag{7}$$

In case both conditions of (a) and (b) happen, the company with a larger variance of BCR_i from the average will be considered first. In this case, if the variance of (a) and (b) are equal, condition (a) will be applied. CT_i can be deducted from CM_i, OCM_i, or a combination of them.

3.3 Game Model of MUT Cost-Sharing

This paper considers financing choice only for the construction phase of MUT projects. In addition, for simplicity, the proposed game model of cost-sharing is based on the allocation of the total construction cost of MUT (CM_t) to only two stakeholders, namely a municipality and an electricity company as shown in Eq. 8.

$$CM_t = CM_m + CM_e \tag{8}$$

where CM_m and CM_e are the MUT construction cost of the municipality and the electricity company, respectively.

Based on financial situation and strategy, both stakeholders have the choice to pay a portion (α_m for the municipality and α_e for the electricity company) of their allocated MUT construction cost share (CM_m or CM_e) from their own fund during the MUT construction phase, and then pay the rest ($1 - \alpha_m$ for the municipality and $1 - \alpha_e$ for the electricity company) by loans or bonds during the MUT operation phase. The municipality can finance by selling bonds and pay back the bonds by annuity payments (a_m) and the electricity company can borrow from financial institutions and pay back the loan by annuity payments (a_e) (Eq. 9) [24].

$$a_i = (1 - \alpha_i)CM_i \frac{r_i}{1 - \left(1 + r_i\right)^{-t_i}} \tag{9}$$

where a_i, r_i and t_i are the annuity payments, interest rate and the payback period (years), respectively, of the bonds/loan for the utility company i.

The interest amount of I_i, as the cost of financing for the utility company i, can be calculated by Eq. 10.

$$I_i = t_i a_i - (1 - \alpha_i)CM_i \tag{10}$$

The game is based on different scenarios of sharing or not sharing the full amount of initial construction cost of MUT by utility companies. It is assumed that a utility company i has the choice to pay its full allocated cost before a due date or to finance it in form of loans (for electricity company) or bonds (for the municipality), and pay installments including principal and interest amounts. The company who is paying fully should be rewarded by the company who is not paying fully [21]. Therefore, there are three scenarios as follows:

Scenario 1: Both companies accept to pay their full allocated portion of the initial construction cost of MUT (i.e. CM_m and CM_e) from their own fund.

Scenario 2: One company accepts to pay its full allocated portion of the initial construction cost of MUT from its own fund. The other company who does not fully pay from its own fund and gets loan/bonds, should pay two extra fees: (1) an interest as the cost of financing I_m/I_e, and (2) a percentage of CM_t, ω, as a reward R (Eq. 11) to the other company.

Scenario 3: Both companies do not pay their full allocated portion of the initial construction cost of MUT from their own fund. Therefore, each company should get loan/bonds and pay the extra fee of interest I_m/I_e.

$$R = \omega CM_t \tag{11}$$

Table 1 shows the game model for MUT construction cost sharing between the municipality and the electricity company.

Table 1. Game model for MUT construction cost sharing

		Electricity company	
		Sharing fully ($\alpha_e = 1$)	Partial sharing ($0 \leq \alpha_e < 1$)
Municipality (water, sewer, and gas network)	Sharing fully ($\alpha_m = 1$)	CM_m, CM_e	$CM_m - R$, $CM_e + I_e + R$
	Partial sharing ($0 \leq \alpha_m < 1$)	$CM_m + I_m + R$, $CM_e - R$	$CM_m + I_m$, $CM_e + I_e$

4 Case Study

The hypothetical case study is a 1 km length MUT to accommodate utility networks of a municipality, including water, sewer, and gas networks, and the electricity distribution network of an electricity company. The following assumptions are considered for the project:

- MUT lifecycle: 2 years for construction and 98 years for operation.
- Design and construction cost of electricity network *(CB_e)* and municipal networks (CB_m) in the buried utilities method are 0.7 and 1.5 M$/km, respectively.
- Total design and construction cost of MUT (CM_t) is 8.8 M$/km.
- The occupied volume of electricity network *(OV_e)* and municipal networks (OV_m) in the MUT are 500 and 1037 m^3/km, respectively.
- Total operation cost of MUT (OCM_t) is 88.2 M$/km.
- The total cost of risk of fire $CR_{fire/design,construction}$ in design and construction phase (safety equipment) is 0.02 M$/km.
- The total cost of risk of fire $CR_{fire/operation}$ in operation phase (insurance) is 0.98 M $/km.
- The anticipated benefit for the electricity company (MB_e) and for the municipality (MB_m) during operation are 14.7 and 49 M$/km, respectively.
- The threshold T for variance of BCR_i from the BCR_{ave} is 15%.
- The assumptions regarding the game theory are: $\alpha_m = 0.3$, $r_m = 10\%$, $t_m = 15$, $\alpha_e = 0.1$, $r_e = 12\%$, $t_e = 30$.
- The assumed value for ω is 6%.

4.1 MUT Cost Allocation Method and Weighted Factors Adjustment

PBC and PUVO methods are selected for the MUT cost allocation in design and construction phases and the operation phase, respectively. The identified risk for adjustment purpose is the risk of fire from gas leakage. Although the risk of fire is caused by accommodating gas and electricity networks in the confined space of MUT, the gas leakage is considered as the main reason for fire. Therefore, the fire risk cost indicator assigned to the municipality (γ_m) and for the electricity company (γ_e) are 0.9 and 0.1, respectively.

The average MUT lifecycle benefit of the municipality networks is larger than that of the electricity company because municipality installs three utility networks compared to one network of the electricity company.

Using the PBC method for design and construction phase (Eq. 1), the values of CM_e and CM_m are 2.8 and 6 M\$/km, respectively. Similarly, using the PUVO method for the operation phase (Eq. 2), the values of OCM_e and OCM_m are 28.6923 and 59.5077 M\$/km, respectively.

Using Eq. 3 for the risk cost adjustment, $CR_{e/fire/design,construction}$, $CR_{m/fire/design,construction}$, $CR_{e/fire/operation}$, and $CR_{m/fire/operation}$ are 0.002, 0.018, 0.098, and 0.882 M\$/km, respectively. The values of CM_e and CM_m after risk adjustment are 2.7956 and 6.0044 \$/km/year, respectively. Similarly, the values of OCM_e and OCM_m after risk adjustment are 28.4715 and 59.7285 M\$/km, respectively.

Using Eq. 4 for the benefit-cost adjustment, the values of BCR_e and BCR_m are 47% and 74.5%. Using Eq. 5 and considering BCR_{ave} of 60.78%, the variance of BCR_e and BCR_m from the average (BCR_{ave}) are both 13.76%, which is less than the assumed threshold T of 15%. Therefore, no benefit-cost adjustment is needed.

4.2 Game Model of MUT Cost-Sharing

Using the adjusted value of CM_m from Sect. 4.1 in Eqs. 9 and 10, the values of a_m and I_m are 0.5526 and 4.086 M\$/km, respectively. Similarly, using the adjusted value of CM_e in Eqs. 9 and 10, the values of a_e and I_e are 0.3123 and 6.854 M\$/km, respectively. Using Eq. 11, the value of R equals to 0.528 M\$/km.

A comparison between the total lifecycle benefit of MUT for the electricity company MB_e = 14.7 M\$/km and the cost of financing plus reward (i.e. 6.854 and 0.528 M\$/km), shows that almost half of the MUT benefit is lost by partial sharing of MUT construction cost. The same comparison for the municipality with the total lifecycle benefit of MUT MB_m = 49 M\$/km and the cost of financing plus reward (i.e. 4.086 and 0.528 M\$/km), shows that less than 10% of the MUT benefit is lost by the municipality. The results of the game model in the construction phase of MUT are presented in Table 2.

A sensitivity analysis can clarify the effects of the input variations on the lifecycle *BCR* of MUT. For example, a combination of the high amount of loan/bond with high-interest rate or/and long payback period can significantly reduce *BCR*. The effect of the cost of risks on *BCR* of MUT may be underestimated. However, a combination of a high number of risks with a high value of risk indicator for a company can reduce *BCR* considerably.

An example of the sensitivity analysis of BCR_e to the loan, and the change of (a) amount of loan ($1 - \alpha_e$), (b) interest rate (r_e), and payback period (t_e) is given in Table 3 and Fig. 5. Nine scenarios are compared based on BCR_e. Scenario 1 is the non-loan scenario of the electricity company ($1 - \alpha_e = 0$). In this case, as indicated in Sect. 4.1, BCR_e is equal to 47%. In Scenario 2, the electricity company gets a loan with interest rate of 12% and a payback period of 30 years, to finance 90% of its allocated MUT construction cost share (CM_e) ($1 - \alpha_e = 90\%$), but the municipality does not sell bonds ($1 - \alpha_m = 0$). Consequently, BCR_e is reduced by 9% after adding the interest of loan I_e and reward R. In the other scenarios, with an individual or a combinational

Table 2. Results of applying the game model (values in M$/km)

		Electricity company	
		Sharing fully ($\alpha_e = 1$)	partial sharing ($\alpha_e = 0.1$)
Municipality (water, sewer, and gas network)	Sharing fully ($\alpha_m = 1$)	6.0044, 2.7956	5.4763, 10.1782
	Partial sharing ($\alpha_m = 0.3$)	10.6182, 2.2676	10.0902, 9.6502

increase of $1 - \alpha_e$, r_e, and t_e, BCR_e is reduced. The largest BCR_e reduction (16.6%) occurs with the combination of $1 - \alpha_e = 95\%$, $r_e = 18\%$, and $t_e = 40$ years (Scenario 9).

Table 3. Sensitivity analysis of BCR_e

Scenario	$1 - \alpha_e$ (%)	r_e (%)	t_e (year)	BCR_e (%)
1	0	N/A	N/A	47.0
2	90	12	30	38.0
3	90	12	40	35.4
4	90	18	30	34.2
5	90	18	40	31.0
6	95	12	30	37.6
7	95	12	40	34.9
8	95	18	30	33.7
9	95	18	40	30.4

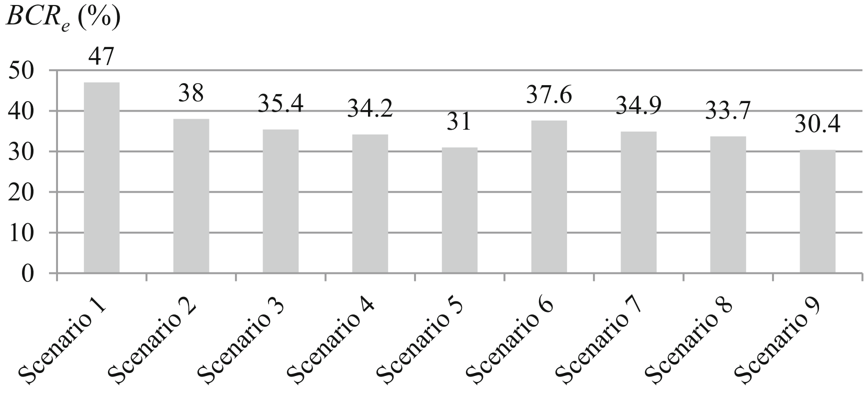


Fig. 5. Sensitivity analysis of BCR_e

5 Summary and Conclusions

This paper proposed a methodology for cost-sharing of MUT projects. In the first step, a cost allocation method must be chosen for the project phases. This paper selected the proportion of buried cost method for design and construction phases and the proportion of utility volume occupancy method for the operation phase of the MUT project. Then, the cost adjustment factors for risk(s) should be identified and the values of risk indicators should be assigned to the utility companies. The allocated costs in the first step must be adjusted by the updated cost of risk(s). For benefit-cost adjustment, the MUT benefit-cost ratios for utility companies should be calculated. The variance between these ratios and the average value should be less or equal to an agreed upon threshold. Otherwise, a portion of the cost of the utility company with the lowest ratio should be transferred to the utility company with the highest ratio to satisfy the threshold condition. In the third step, based on the financial situation and strategy, the utility companies have the choice to pay a portion of their adjusted MUT construction cost share from their own fund, and the rest by loans/bonds. The municipality can finance by selling bonds and pay back the bonds by annuity payments. Other utility companies can borrow from financial institutions and pay back the loans by annuity payments. If loans/bonds financing is chosen, a reward must be paid to the other company who is paying fully, and the total interest, as the cost of financing, must be calculated and added to the reward. For the company that paid the MUT full allocated costs from their own fund, *BCR* equals the lifecycle benefit of MUT plus the reward (if applicable) divided by the allocated lifecycle cost. For the company that used loan/bond, *BCR* equals the lifecycle benefit of MUT divided by the sum of the allocated lifecycle cost of MUT, the interest of loan/bond, and the reward (if applicable). The goal of each company is maximizing its *BCR*. Furthermore, the estimated *BCR* can be considered as an important indicator for utility companies to decide on participation in a MUT project. The contributions of the proposed method of this paper include: (1) considering game theory with loan/bond option for MUT stakeholders to finance the project fully or partially from financial institutions; (2) improving fairness of MUT cost allocation by establishing cost adjustment factor of risk; (3) improving fairness of MUT cost allocation by establishing cost adjustment factor of *BCR*; (4) incorporating MUT cost allocation and adjustment, financing, and reward methods within game theory to reach an optimal cost allocation amongst the stakeholders. Future research should be focused on the following: (1) identifying more risk adjustment factors; (2) applying game theory for lifecycle cost-sharing of multiple utility companies; (3) developing methods for the calculation of reward of utility companies; (4) applying sensitivity analysis to investigate the effect of the most important input variables, such as interest rate, payback period, and risks, on the lifecycle *BCR* of MUT; and (5) comparing total lifecycle cost of MUTs with traditional method of buried utilities by considering quantified social, environmental, and health costs.

References

1. Yang, C., Peng, F.L.: Discussion on the development of underground utility tunnels in China. Procedia Eng. **165**, 540–548 (2016)
2. Hunt, D., Nash, D., Rogers, C.: Sustainable utility placement via multi-utility tunnels. Tunn. Undergr. Space Technol. **39**, 15–26 (2014)
3. Cano-Hurtado, J.J., Canto-Perello, J.: Sustainable development of urban underground space for utilities. Tunn. Undergr. Space Technol. **14**(3), 335–340 (1999)
4. Laistner, A.: Utility tunnels long-term investment or short-term expense? The new economic feasibility of an old idea. In: 15th International No Dig Conference, Taipei (1997)

5. Rogers, C., Hunt, D.: Sustainable utility infrastructure via multi-utility tunnels. In: 1st International Construction Specialty Conference, Calgary (2006)
6. Canto-Perello, J., Curiel-Esparza, J.: Assessing governance issues of urban utility tunnels. Tunn. Undergr. Space Technol. **33**, 82–87 (2013)
7. Gilchrist, A., Allouche, E.N.: Quantification of social costs associated with construction projects: state-of-the-art review. Tunn. Undergr. Space Technol. **20**(1), 89–104 (2005)
8. Najafi, M., Kim, K.O.: Life-cycle-cost comparison of trenchless and conventional open-cut pipeline construction projects. In: Pipeline Division Specialty Congress 2004, San Diego (2004)
9. Cle de Sol: Guide practique des galeries multireseaux. Clé de Sol, Voiron Cedex: Éditions Techni.Cités (2005)
10. Canto-Perello, J., Curiel-Esparza, J.: Risks and potential hazards in utility tunnels for urban areas. Proc. Inst. Civ. Eng. Municipal Eng. **156**(1), 51–56 (2003)
11. Kang, Y.K., Choi, I.C.: Economic feasibility of common utility tunnel based on cost-benefit analysis. J. Korean Soc. Saf. **30**(5), 29–36 (2015)
12. Ormsby, C.: A framework for estimating the total cost of buried municipal infrastructure renewal projects. M.S. Thesis, Department of Civil Engineering and Applied Mechanics, McGill University, Montreal (2009)
13. De Marcellis-Warin, N., Peigner, I., Mouchikhine, V., Mahfouf, M.: Évaluation des coûts socio-économiques reliés aux bris d'infrastructures souterraines au Québec. CIRANO, Montréal (2013)
14. Sterling, R., Admiraal, H., Bobylev, N., Parker, H., Godard, J.-P., Vähäaho, I., Rogers, C., Shi X., Hanamura, T.: Sustainability issues for underground space in urban areas. Proc. Inst. Civ. Eng. - Urban Des. Plann. **165**(4), 241–254 (2012)
15. CERIU: Guide pour l'évaluation des coûts socio-économiques des travaux de renouvellement des conduites d'eau potable et d'égout. Gouvernement du Québec, MAMROT, Montréal (2010)
16. Canto-Perello, J., Curiel-Esparza, J., Calvo, V.: Analysing utility tunnels and highway network coordination dilemma. Tunn. Undergr. Space Technol. **24**(2), 185–189 (2009)
17. Manuilova, A., Dormuth, D.W., Vanier, D.J.: A Case Study of Use and External Components of Social Costs that are Related to Municipal Infrastructure Rehabilitation, National Research Council Canada (NRCC) (2009)
18. Hunt, D., Rogers, C.: Barriers to sustainable infrastructure in urban regeneration. Proc. Inst. Civ. Eng. - Eng. Sustain. **158**(2), 67–81 (2005)
19. European Commission: Guidelines for Successful Public-Private Partnerships (2003). http://ec.europa.eu/regional_policy/sources/docgener/guides/ppp_en.pdf. Accessed 21 Dec 2017
20. CPAMI: Regulations on Cost Allotment of Common Duct Construction and Management. Construction and Planning Agency, Ministry of Interior, Taiwan (2011). https://www.cpami.gov.tw/public-information/laws-regulations/10-public-works/10777-regulations-on-cost-allotment-of-common-duct-construction-and-management.html. Accessed 10 Jan 2018
21. Gui, X., Wang, W., Zhang, S.: The incentive mechanism for financing of the municipal utility tunnel construction. Chin. J. Undergr. Space Eng. **7**(4), 633–637 (2011)
22. Myerson, R.B.: Game Theory: Analysis of Conflict. Harvard University Press, Cambridge (1991)
23. Wang, X., Zhu, F.: Research of charge strategy for urban municipal utility tunnel based on game theory analysis. Chin. J. Underg. Space Eng. **9**(1), 197–203 (2013)
24. Finance Formulas. http://financeformulas.net/Annuity_Payment_Formula.html. Accessed 22 Dec 2017

Data-Driven, Multi-metric, and Time-Varying (DMT) Building Energy Benchmarking Using Smart Meter Data

Jonathan Roth and Rishee K. Jain(✉)

Urban Informatics Lab, Department of Civil and Environmental Engineering,
Stanford University, Stanford, CA 94305, USA
rishee.jain@stanford.edu

Abstract. New and emerging data streams, from public databases to smart meter infrastructure, contain valuable information that presents an opportunity to develop more robust data-driven models for benchmarking energy use in buildings. In this paper, we propose a new Data-driven, Multi-metric, and Time-varying (DMT) energy benchmarking framework that utilizes these new data streams to benchmark building energy use across multiple metrics at the daily time scale. High fidelity data from smart meters enables the DMT benchmarking framework to produce daily benchmarking scores and use daily weather data to understand seasonally adjusted performance. Intra-day building efficiency is also investigated by benchmarking buildings across several metrics (e.g., total energy usage, operational energy usage, non-operational energy usage) thereby enabling deeper insights into building operations than traditional yearly benchmarking models. By using quantile regression modeling, the DMT framework can differentiate and understand the main drivers of energy consumption between low and high performing buildings and between building operational states. To illustrate the insights that can be gleaned from the proposed DMT framework, we apply the framework to understand building performance for over 500 schools throughout the state of California. The DMT framework provided insights into how various drivers impacted energy usage for both high and low performing buildings, and results indicated that schools had consistent drivers of energy usage. Overall the DMT framework was designed to be highly interpretable such that it could help bridge the gap between data science and engineering methods thus enabling better decision-making in respect to energy efficiency.

Keywords: Building energy · Energy benchmarking · Data-driven
Time-varying · Smart meter

1 Introduction

Rising energy costs and global warming concerns are causing municipalities, utilities, and building owners alike to adopt energy efficiency solutions [1]. Given that buildings account for over 75% of U.S. electricity consumption, effectively identifying inefficient buildings and their sources of waste can be tremendously valuable for the construction of effective energy efficiency policies and incentives [2]. Specifically, the operational

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 568–593, 2018.
https://doi.org/10.1007/978-3-319-91635-4_30

phase of buildings account for 80–90% of its life cycle energy, indicating that a large focus needs to be placed on improving the energy efficiency of existing buildings (and not just the design of new buildings) if we hope to reduce overall energy consumption and its associated negative environmental impacts [3]. However, identifying which buildings to target for potential energy efficiency opportunities, incentives, and programs can be very difficult as it requires discerning inefficient buildings from efficient ones. This problem is further exacerbated by the fact that building energy inefficiency is usually caused by a combination of technical and social dynamics that can be difficult to separate [4]. Historically, engineers have turned to on-site energy audits and building simulation models as the primary avenues to determine potential sources for energy savings within a building. However, both solutions face significant barriers to be deployed at scale as they are capital, labor, and expertise intensive processes.

In the United States and across the world, smart energy meters are quickly being adopted. In fact, there are now over 70 million smart energy meters in the United States [5]. This widespread penetration has brought with it a plethora of high-fidelity time-series energy consumption data. This data is usually captured at the building level and at time resolutions of one hour or smaller and thus provides a unique opportunity to understand the complex energy dynamics of a building. Moreover, the emergence of such new energy data streams is enabling the development of new data-driven approaches to assess building energy efficiency and provide high-level insights into building operations that are less capital and labor intensive.

Governments are taking notice of these new sources of data, and many have begun to utilize data-driven energy benchmarking solutions to isolate poor performing buildings; twenty cities across the U.S. have mandated the use of energy benchmarking and are currently working on creating legislation around their results [6]. The main purpose of benchmarking is to rank a building's operational performance to either a performance baseline or to other buildings, allowing for the identification of inefficient energy users. Benchmarking can be accomplished through a point based system (like LEED), simulation, or data-driven approach. Point based systems either overlook or oversimplify the operation side of buildings, while simulations, though often insightful, are expensive and time-intensive. Data-driven methods are low cost, easy to implement, and can be regularly updated.

Additionally, a main difficulty in energy benchmarking is quantifying the value or utility that building occupants receive from using energy [7]. Occupants value countless services offered by a building and are often unaware of their own preferences [8]; this suggests that efficiency is a relative term that is defined in the context of the services offered by a building. Buildings also have constantly fluctuating occupancy levels and serve a wide variety of people that interact with the building in diverse ways. As a result, it is integral that a building energy benchmarking system maintains interpretability such that it provides building managers and owners the opportunity to assess areas of inefficiency and weigh the tradeoffs between efficiency, costs and utility of their buildings. An interpretable benchmarking system can provide insights that enable owners to make decisions that have synergistic effects, like improving the thermal comfort of its occupants in addition to lowering energy costs. In the end, an effective benchmarking system will empower these decision-makers by providing greater information that will enable them to co-optimize building utility and energy efficiency.

This paper proposes a novel Data-driven, Multi-metric, and Time-varying (DMT) energy benchmarking framework based on quantile regression. The DMT benchmarking framework aims to harnesses new emerging high-fidelity smart meter energy data streams and maintain interpretability such that decision-makers can gain deep insights into operational performance that is currently not possible with traditional benchmarking methods. By utilizing smart meter data, it is uniquely capable of producing scores and information across an array of metrics at the daily temporal scale. Examining energy consumption dynamics at the daily scale allows users to understand performance differences between days and seasons, enabling deeper insights into drivers of energy. To understand intra-day performance, the DMT framework also benchmarks building energy performance on operational and non-operational daily energy use. In practice, the DMT benchmarking framework can provide building managers and utilities with more immediate feedback on performance, thereby taking advantage of behavioral effects that can lead to increased energy savings [9, 10]; the more immediate the feedback, the greater the energy savings and educational effects of the provided information.

2 Literature Review

In order to contextualize our proposed DMT energy benchmarking framework amidst the current literature, we conduct a literature review to examine current energy benchmarking models and highlight their strengths and weaknesses. We also review current work that is beginning to utilize new emerging smart energy data streams to better understand energy consumers. Lastly, we review current methods in fault detection and diagnosis (FDD) and measurement and verification (M&V) to demonstrate the need for new data-driven, performance based approaches.

2.1 Energy Benchmarking and Data-Driven Approaches

The growing availability of data has recently driven the development of new energy benchmarking systems for buildings. Traditionally, limitations in data have constrained benchmarking to the use of a key performance indicator (KPI) due to its simplicity and low cost. Energy use intensity (EUI) is the most common KPI and ranks buildings by normalizing their energy consumption by their area [11]. Despite its ease of use and basic interpretation, KPIs like EUI do not normalize for other important factors in buildings like HVAC systems, buildings' age, occupancy levels, and weather that are known to have effects on energy performance [12].

The most popular energy benchmarking system is EnergyStar, which is based on ordinary least square (OLS) regression that can normalize for a variety of factors and find the average consumption for the set of input buildings [13]. The scores for the buildings are then based on the residuals; however, the residuals encompass statistical noise, unexplained factors, and any measurement error [12]. This means that the regression line poorly models the energy dynamics of buildings with very large and small residuals, indicating poor explanatory power for such buildings. In energy benchmarking, the buildings that consume abnormally high and low amounts of energy (for its given set of inputs) are of greatest interest, yet OLS regression only provides an estimate for the average consumer. Additionally, the buildings at the tail-ends of the distribution can greatly skew the regression line, further diminishing its effectiveness of appropriately benchmarking buildings and identifying efficient or inefficient buildings. The EnergyStar system uses OLS as its benchmarking model but has several additional weaknesses that can lead to erroneous and potentially misleading scores. First, EnergyStar requires human data input which can often be inaccurate or missing [14]. Second, EnergyStar uses a national level database of several thousand buildings to construct its model, meaning it overlooks local effects that have large impacts on building energy consumption. These data limitations and the sensitivity to outliers hinders the ability of EnergyStar and other OLS models to effectively benchmark buildings at a local scale.

Data envelopment analysis (DEA) is a nonparametric empirical model that compares a building with the best-performing buildings of its own class [15]. The model constructs a frontier that represents the minimum amount of energy used across all input variables for the buildings in the dataset [16]. This allows for the computation of necessary improvements in the inputs to make a building efficient [17, 18]. Since DEA is nonparametric, it does not assume any functional form of the frontier, but is often criticized for its deterministic and non-statistical nature [12]. Outlier data is very problematic and can greatly skew the constructed frontier, resulting in many buildings being deemed inefficient. Furthermore, adding factors or variables to the model never decreases efficiency scores and will result in an over estimate of efficient buildings. DEA also falsely assumes a constant rate of returns (i.e., returns to scale), meaning that linearly scaling the inputs will have a linear effect on the state of the output. Finally, the nonparametric nature of DEA inherently requires the model to be reconstructed anytime new data is added, limiting its flexibility in practical applications. To address several of these issues, Kavousian et al. [19] was one of the first to utilize smart meter infrastructure in benchmarking and extend DEA to produce uncertainty estimates for efficiency scores. The high-fidelity smart meter data enabled uncertainty to be calculated for scores by using repeated measurements, rather than assume a deterministic consumption level. However, DEA is still limited in its ability to handle outlier data, is prone to overestimating efficiency when there are a high number of covariates, and assumes a constant rate of return for the input variables.

Stochastic frontier analysis (SFA) is another frontier method that addresses several of the problems with DEA. Most importantly, it separates the error components from the inefficiency components [20]. The level of inefficiency can then be directly quantified, as it is not contained under one umbrella error term. SFA is a parametric method, unlike DEA, where a regression is built under the assumption that the two error components are independently distributed [21]. The inefficiency indicator term is a one-sided non-negative random disturbance term that can vary over time and is assumed to follow a half-normal distribution. The measure of efficiency for every building takes on a value between zero and infinity, where zero is considered the most efficient. However, existing studies have mainly focused on applying SFA to one building type or cluster, as the model is unable to construct a frontier if there is a significant amount of variance in the data [22]. Similar to DEA, the SFA method is sensitive to outliers, is prone to creating an overestimate of the number of efficient buildings, and struggles to handle highly variable data.

Smart meter technology is now enabling more real-time monitoring and tracking of electricity consumption at much smaller time intervals than traditional meters [23, 24]. A considerable amount of research has been done at the residential scale to cluster similar energy consumers and better understand how these customers use their energy [25, 26]. Literature has also examined clustering data separately by season and weekday and weekends due to known differences in usage patterns [25, 27]. Clustering on energy use and other attributes can assist distributed network operators (DNOs) better identify suitable customers for energy management solutions and can enable greater control over grid operations [28]. Clustering techniques can also be used to identify faulty metering, overloaded or ageing components, and irregular behavior due to external unknown factors [29, 30]. Other studies utilizing smart meter data are concerned with identifying suitable customers for demand-side response and diagnosing behavioral patterns associated with different types of energy feedback. (building type, socio-demographics, etc.) [31–33]. The proposed DMT benchmarking framework aims to utilize this new data stream to parameterize daily energy use profiles to enable better identification of inefficient buildings, demand response customers, and changes in energy use patterns.

2.2 Fault Detection and Diagnosis (FDD)

Building energy consumption is driven by a dynamic and complex host of factors that can be difficult to understand. Teasing out normal variations in energy consumption driven by occupancy and weather effects from building system degradation or improper operations is crucial for energy conservation. These abnormal energy consumption faults are difficult to monitor and diagnose, and typically focus on building system equipment like HVAC systems or air-handling units [34, 35]. Most research on fault detection and diagnosis focus on subsystem and building system equipment due to their unit specific and highly granular data streams; sensors for building automation systems (BASs) typically collect data such as electric power, temperature, humidity, flow rate, pressure, and CO2 concentration levels [36]. Using these types of data streams, research has shown that data-driven approaches using machine learning algorithms, from support vector machines (SVMs) to neural networks, can accurately isolate and detect faults within these subsystems [37, 38].

However, there has been a lack of research focusing on FDD for whole building energy consumption using just smart meter infrastructure or whole building energy consumption, specifically at greater scales across a large portfolio of buildings [39]. This is often referred to as whole building level diagnosis, which does not necessitate as much information on building operations as most research concerning FDD requires [40, 41]. Though this data stream is not as information rich as the streams of data pouring in from BASs, smart meter infrastructure is much more ubiquitous, thus theoretically allowing for FDD to be administered over larger, urban scales. Currently, utilities often spend a disproportionate amount of resources constructing relationships with their largest consumers, and only reach other customers through mass-marketing solicitation approaches [42]. Providing data-driven benchmarked scores on several metrics will help utilities better understand power demand from their customers and better target buildings with specific retrofits or promotions. Utilities could better coordinate with municipalities and ESCOs (energy service companies) to target buildings, decrease customer acquisition costs, and increase investment potential from ESPCs (energy-saving performance contracts) [43]. In order to extend FDD methods and utilize new energy data streams, our proposed DMT benchmarking framework aims to produce performance-based metrics that can be used to identify inefficient operations in buildings and be used as input data for future FDD algorithms.

2.3 Measurement and Verification (M&V)

Traditionally, measurement and verification (M&V) has been conducted using physics and engineering based approaches, largely due to limitations in M&V budgets for data collection [44, 45]. Alternative approaches have focused on building simulation models to reduce error and achieve more accurate savings estimates [46]. Recent research and industry standards have adopted data-driven approaches that create a baseline model on pre-retrofit energy data to predict future consumption. This prediction can then be used to compare against observed energy consumption post-retrofit, where the savings is the difference between the two [47]. These methodologies are susceptible to three types of uncertainty that can cause inaccurate results: model uncertainty, input/output measurement uncertainty, and sampling uncertainty [48]. New models are attempting to reduce these issues by accounting for uncertain input data in the learning of model hyper-parameters [49]. These data-driven approaches have mostly focused on predicting future consumption for buildings using various techniques, but have forsaken performance based metrics. Our proposed DMT energy benchmarking framework aims to provide further insight for M&V application by establishing multi-metric, time-varying scores that utilize public data sources and normalize for performance over a large portfolio of buildings. Rather than predict total savings, this strategy can help decision makers understand the improvements in efficiency from retrofits relative to other buildings in a portfolio and aid in deciding which buildings to target for efficiency improvements.

3 Methodology

The primary objective of this paper is to propose a novel a Data-driven, Multi-metric, and Time-varying (DMT) framework for benchmarking building energy usage. The DMT methodology aims to leverage emerging data streams from smart meters to provide insight into how buildings consume energy and provide metrics on relative levels of energy efficiency. As a result, the DMT framework extends previous work that utilizes quantile regression [50] to establish new multi-metric and time-varying benchmark scores. We chose to utilize quantile regression as a basis for our model as it addresses several key issues with other current leading benchmarking practices. Specifically, quantile regression is able to: (a) reduce sensitivity to outliers; (b) handle heteroskedastic data; (c) normalize for numerous explanatory variables; (d) provide a theoretical maximum level of performance; (e) eliminate an over-estimation of efficient buildings when more covariates are included; (f) identify non-linear relationships between explanatory variables and consumption; and (g) model the entire conditional distribution of the dependent variable. The multi-metric and time-varying benchmark scores enable finer and deeper insights into building performance to a wide array of stakeholders including building owners, operators, and policy-makers.

In this section, we provide a brief overview of the mechanics of quantile regression, our proposed extension to the multi-metric and time varying cases, and finally the insights that can be gleaned from the outputs of the quantile regression modeling process.

3.1 Quantile Regression

Determining benchmarking scores for buildings using quantile regression requires 2 main steps: (1) Construct a suite of quantile regression models for a range of quantiles (tau); (2) Assign scores for each building based on the closest model prediction.

Quantile regression models the quantiles (tau values) of the conditional distribution of the response variable as functions of observed covariates [51]. It is especially useful when outliers need to be examined, as is the case for benchmarking, which aims to identify the best and worst energy performing buildings [52]. Quantile regression takes on the same function form as Ordinary Least Squares (OLS) regression, as shown in Eq. (1).

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \tag{1}$$

However, the loss function for quantile regression relies on the sum of absolute deviations instead of the sum of squares errors (like in OLS). The least absolute deviation (LAD) can then be written as the objective of the cost function of quantile regression, as seen in Eq. (2). Using LAD as the loss function downplays the importance of outliers and allows for exploration into other quantiles by minimizing the sum of absolute residuals asymmetrically and applying a set weight,

$$\mathrm{Q}(\beta_\tau) = \arg\min \sum_{\mathrm{i}=1}^{\mathrm{N}} \rho_\tau\left(\mathrm{y}_\mathrm{i} - \mathrm{x}_\mathrm{i}'\beta_\tau\right) \tag{2}$$

where ρ_τ serves as a check function as shown in Eq. (3):

$$\rho_\tau(\mathrm{x}) = \begin{cases} \tau * \mathrm{x}, & \text{if } x \geq 0 \\ (\tau - 1) * \mathrm{x}, & \text{if } x < 0 \end{cases} \tag{3}$$

Here, τ (tau) is the sample quantile, taking on a value between 0 and 1, where a value of 0.5 corresponds to the median. N is the total number of data points, y_i is the response variable, x_i' is the vector of covariates, and β_τ is the produced vector of coefficients for the given value of τ. By setting tau to values between 0 and 1, quantile regression can model other quantiles of the dataset and provide a more complete picture of the relationship between the covariates and dependent variable, which is especially important for data with heterogeneous variances like building data.

A major concern in the application of OLS in energy benchmarking is the heteroskedastic nature of building data, as this can invalidate statistical tests that assume modeling errors are uncorrelated and uniform. In the presence of heteroskedasticity, the Gauss-Markov theorem is not applicable, meaning that OLS estimators are not the Best Linear Unbiased Estimators and their variances are not the lowest of all unbiased estimators [53]. This can lead to biased standard errors that can cause hypothesis tests to be incorrect and result in coefficients that will be missing information on the relationship of the estimators in subpopulations that have different variability from the others. Quantile regression avoids these pitfalls and can handle heteroskedastic data and model the entire conditional distribution, which enables its application to more varied types of datasets as it does not require normal residuals or constant variance. Quantile regression therefore exposes relationships between predictors and the response variable in other parts of the distribution, not just the median, and measures their rates of change and relative importance.

To produce benchmarked scores, models are created at each tau value (0.01 to 0.99) to model building consumption across the conditional distribution. The resulting 99 models produce 99 separate predictions for any given set of inputs (building and weather characteristics – area, enrollment, mean temperature, etc.). The score for a building is then determined using Eq. (4),

$$\text{score} = (1 - \tau_{closest}) * 100 \tag{4}$$

where $\tau_{closest}$ is the tau value of the model with the closest predicted consumption to the building's observed energy consumption. For example, a building that consumed 100 kWh for the day may be assigned a score of 20 because the predicted energy consumption for the quantile regression model with a tau value of 0.80 is closer to 100 kWh compared to the other 98 models. Therefore, the building falls into the 20th percentile of energy consumption and receives a score of 20, since $(1 - 0.8)$ *100 is equal to 20, meaning that 20% of buildings underperformed this building (a score of 99 is the best attainable score as a tau value of 0.01 is the smallest tau used in this example). This scoring system is comparable to EnergyStar where higher scores translate to better performing buildings.

For the purposes of using quantile regression in energy benchmarking, a variable selection process was constructed with the aim of maintaining the same set of variables for each quantile, so that resulting models are more interpretable. Otherwise, a different subset of variables may be selected for each model, resulting in coefficient values of zero for unselected variables; consequently, valuable insights would be lost (Sect. 3.3 discusses how to interpret the output of these models). In order to counter this issue, a modified forward variable selection process was constructed, where each quantile model is initialized with zero starting variables, and the *simultaneous addition of each independent variable for all quantile models* is then examined. The addition of the variable that minimized the sum of the cost functions *across all models* is then selected to add to the model. This process is then repeated, testing the addition of each variable on all the models, until a cutoff threshold is met.

3.2 Extension to Multi-metric and Time-Varying Scores

We propose extending quantile regression to model consumption metrics at the daily scale using smart meter data. This provides a daily relative performance across several parameters which in turn enables building managers to more frequently compare their performance to their peers and identify abnormal consumption patterns and operational paradigms. Moreover, utilities can also utilize the daily scores to examine differences across a portfolio of buildings and help identify potential customers for certain types of energy efficiency retrofits and rate plans. Daily benchmarking allows for examination into seasonal energy efficiency patterns which can be used to recognize buildings that are sensitive to temperature variations. This study explores three different metrics for daily benchmarking: *total daily energy consumption*, *energy consumption during operating state*, and *energy consumption for non-operating state*. These scores in isolation do not provide information on potential improper operations but rather can be used as comparative metrics for understanding relative performance of a building over time.

In this study, we utilize a simple heuristic for denoting the operational period as our building sample is comprised of school buildings that have established schedules (see Sect. 4 for more information of the data). The operational state is designated as energy consumed between 7 am to 3 pm while the non-operational state is energy consumed between 9 pm to 4 am. We note that our proposed DMT benchmarking framework can utilize emerging techniques for clustering smart meter data [25] to determine the operational period, but we have decided to select set time-periods for partitioning due to known operating schedules of school buildings and to simplify this process in this initial study. Future work aims to extend clustering techniques to automatically distinguish between operating states. Consumption levels during these designated times represent two drastically different states for the school buildings. The operating state (7 am–3 pm) is when the school building systems are fully functioning to meet the demands of its occupants. In contrast, the non-operating state (9 pm–4 am) is during a time when building systems should be at low or zero functioning capacity since this time period is during the middle of the night when school buildings have low occupancy levels.

Daily benchmarking also allows for the inclusion of other information rich, high-fidelity data streams. Daily weather data can be used as input parameters for this type of modeling, thereby greatly enhancing the amount of information captured when compared to yearly weather data. Traditional yearly benchmarking models overlook this dynamic relationship because it is limited to yearly weather metrics, which can lead to misleading results, since weather has a huge impact on building energy consumption [54]. Furthermore, data at the daily scale allows for finer tuning when deciding which, and how much historical data to include in the modeling process. Effects of inherent temporal variations in energy consumption throughout the year can be captured by limiting the included data to monthly or seasonal scales. In this paper, we model each month independently for each of the three metrics. For example, daily benchmarked scores for the month of June are determined by only including building data from that month. This partitioning of the data results in benchmarked scores for each building for each month, where scores can then be aggregated for the entire year for comparison. The proposed DMT benchmarking framework is adaptable to work with different choices of partitioning methods; however, we chose to partition buildings on a monthly basis as it enables us to (1) cluster building energy consumption in similar temporal and seasonal buckets to make for more equitable comparisons; (2) examine the efficacy of the variable selection process by examining the selected variables across months; (3) reduce computational burden for computing models as only 12 models need to be constructed per year. This approach also lends itself for easier and automatic recalibration, unlike EnergyStar, since it does not necessitate an entire year's worth of data. Instead, the model can be constructed using data from the most recent month rather than require a manual update to a national database of buildings using yearly data.

Once the scores for all three metrics are calculated for each building, then individual school buildings can be examined to derive insights into potential sources of poor performance. Scatter plots for total daily consumption scores can be constructed for each individual building to observe seasonal trends and identify abnormal days of consumption. For example, a building may consistently receive daily scores between 70 and 80 but receives a score of 30 during one weekday. This may be caused by a large increase in occupancy, clogged filters in the HVAC system, or a failure in the control system to shut down systems during nighttime (non-occupational state). Examining a large increase in the non-occupational state score for the same day may suggest that the issue was a control problem rather than a clogged filter. Seasonal decomposition may also show that certain buildings perform much worse during summer months. This may suggest that the building is sensitive to warmer temperatures, indicating potential issues with its AC unit.

In addition to aiding in diagnosing daily building system mishaps, the DMT benchmarking framework can be used to help detect much larger, systemic issues. A control system that frequently fails to shut down during nighttime hours can have a substantial impact on the energy consumed in a building. The DMT benchmarking framework can aid in finding these systems across a portfolio by identifying buildings that have the largest difference between average operational and non-operational scores throughout an entire year. This difference represents buildings that perform well during operating hours but use substantial amount of energy compared to their peers during non-operating hours. These differences can also be independently examined for weekday and weekend days to separate effects that may occur for days with lower building occupancy.

3.3 Energy Drivers and Interpretation

The use of quantile regression for multi-metric and time-varying benchmarking enables insight into drivers of energy consumption across three-dimensions. First, each suite of quantile regression models produces a set of coefficient values, corresponding to each independent variable and each tau value. These values provide valuable insights into the changing effects of the variables throughout the conditional distribution (i.e., differences in drivers for low and high performing buildings). Second, when the data is partitioned by month, to separately create scores for each month and for each metric, the effect of energy drivers can be examined over time while adjusting for seasonal differences. This produces daily scores using only data from the same month but gives different fitted values for the covariates for each month. Third, benchmarked scores for each metric produces its own set of coefficients for each driver. This allows for comparison of energy drivers between benchmarked metrics (i.e., operational states).

First and foremost, the energy drivers for the entire condition distribution can be determined through quantile regression as each tau value corresponds to a different quantile. A set of influence plots for each independent variable is created for each month and for each metric, since these are modelled separately. Figure 1 below is an example influence plot showing the learnt coefficients (β) for the cooling degree day (CDD) variable, and its effect on total daily energy consumption, associated with each constructed quantile model for the month of June. The black dotted line in the plot reveals that CDD is between two and four times as impactful for low performing buildings compared to efficient buildings for total daily energy use in the month of June (gray shaded area indicates 90% confidence interval). This is determined by comparing the coefficient values between the low and high tau values, which correspond to high and low scores, respectively (as discussed in Sect. 3.1). If only OLS was used for benchmarking, the varying effects of CDD on different parts of the distribution would have gone unnoticed, as shown by the single red solid line representing the value obtained by OLS (red dashed lines are 90% confidence interval).

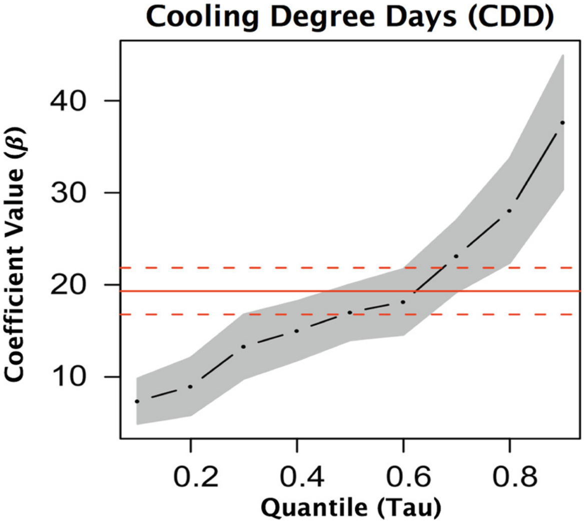


Fig. 1. Influence plot showing the varying effect of cooling degree days for total daily energy consumption for June across all the quantiles. (Color figure online)

Second, by modeling each month independently, influence plots can be constructed for each metric for each month. This can reveal changing effects of variables over time and between seasons. Continuing the example above, the influence plot for CDD for the month of October may have a shallow slope and only range from 10 to 20. This would indicate that poor performing buildings are not as greatly affected by CDD in November as they are in June. Several hypotheses can be tested and later validated to determine this result. For example, as hot days are much more common in June than November, low performing buildings may have AC units that struggle to condition the building during consecutive hot days; or the thermal mass of these low performing buildings helps condition the space during the occasional hot day in November.

Lastly, comparing influence plots for the same independent variable but between benchmarked metrics may also reveal insights into building operations. The influence plots resulting from the operating state metric for CDD for the month of June may have a similar shape to that shown in Fig. 1 despite only measuring energy consumption for a fraction of the day. Similar influence plots for both the total daily energy consumption and the operation state would suggest that the building uses close to no energy during non-operating hours. This could then be confirmed by examining the influence plots for the non-operation state scores.

These various comparisons can inform building managers, energy policymakers, and utilities about the key drivers of energy consumption across operating states, over time, and between high and low performers. Such insights can reveal potential focal points for building retrofits and help decision-makers decide how to best tailor their policies and allocate financial and non-financial incentives for improved building operations.

4 Data

The efficacy of the proposed DMT framework is demonstrated on smart meter data released from Proposition 39 (Prop39) through the California Energy Commission (CEC). This dataset contains smart meter readings for over 4,000 schools throughout the entire state of California from July 2012 until July 2015. This dataset was combined with daily weather data from the National Oceanic and Atmospheric Administration (NOAA) at the zip code level. To get more information on school building characteristics and demographics, data from the California Department of Education (CDE) and the Federal Census Bureau was supplemented to the dataset.

The Prop39 dataset was selected to examine the DMT benchmarking framework for several reasons. First, California has allotted nearly $550 million annually for five years towards improving the energy efficiency of its school buildings. However, the state is having difficulty in allocating these funds to achieve maximum energy savings and improvements in facility operations; an energy benchmarking framework that utilizes smart meter data streams can aid in better fiscal distribution and could enhance the efficacy of energy efficiency incentive dollars. Second, a school dataset is advantageous from a methodological perspective as it provides a relatively homogenous basis for testing our DMT benchmarking framework. In other words, it is a reasonable assumption that all schools are engaged in relatively similar activities, have similar operating schedules, and are subject to the same educational trends throughout the state of California, and thus provide a good baseline dataset for testing. Finally, this school building dataset allows for supplemental data to be scraped from public sources, like the CDE, Census Bureau, and CEC that would not otherwise be readily available for private buildings, and provides a more holistic understanding of building energy dynamics.

After cleansing and removing erroneous data, the total number of buildings in the dataset was reduced to 569 buildings. For the purposes of this study, the daily performance of school buildings throughout the 2014–2015 (July 2014 to June 2015) academic year was investigated. This timeframe allows for a full twelve months to be examined, enabling investigation into seasonal and operational differences. Predictive mean matching was used to impute any missing values [55].

5 Results and Discussion

Using the forward variable selection process described in Sect. 3.1, quantile regression models were built for each metric (total daily energy, operational energy, non-operational energy) and for each month during the 2014–2015 academic year for a total of 36 models. We decided to use the variable selection process to determine the top 10 energy drivers for each month, rather than choose a cutoff criterion with a minimum reduction in summed cost function values. Each model examined all tau values between 0.05 and 0.95 (with a stepsize of 0.01), where the extreme quantiles were excluded because of large standard errors. Due to the variable selection process, each of the 36 models independently selected the variables that it found to be most important. This process was utilized because building energy drivers during summer months are known to be very different than those during the winter months. Therefore, our model can capture these differences by including a different subset of variables for each month which enables seasonally adjusted benchmarking scores for all the buildings. Table 1 below shows the most frequently occurring variables selected across all 12 months for each of the three-different metrics. The table only includes those variables that were selected more than 50% of the time (6 of 12 months) for at least one of the three benchmarked metrics. Information and descriptions about each of the variables used in the study can be found in Appendix A.

Results of the variable selection process show that, regardless of season, several variables are consistent drivers of total energy use: area, enrollment, school type, charter school, and high grade. The indicator variable for weekend showed up in every model for both total daily energy and operation state metrics, yet only occurred in the non-operation state model for one of the twelve months. Intuitively, nighttime energy use (non-operational state) for school buildings will have similar energy consumption levels whether it is during a weekday or weekend. The same does not hold true for daytime hours (operation state) where schools are being highly utilized during weekdays and have low to no occupants during weekends. Therefore, the models for the non-operational state determined that the inclusion of the weekend variable did not add explanatory power to the final model. Additionally, the non-operational state included a more varied set of variables for each month, with both median household income and enrollment per area frequently being selected. The median household income variable is similar to the current expense ADA (Average Daily Attendance – see Appendix A) variable which acts as a proxy for school wealth and the state of the building facilities. Including median household income and enrollment per area indicates that these two variables add more explanatory power to the model than other excluded variables, despite the presence of current expense ADA and enrollment.

Table 1. Frequency of selected variables for each benchmarked metric. Only those variables that were selected greater than 50% of the time for at least one benchmarked metric are shown. * Temperature variable includes the occurrence of at least one of: mean temperature, mean temperature squared, cooling degree day (CDD), cooling degree day squared (CDD_2). ** Free or Reduced Priced Meals variable includes the occurrence of at least one of: Free or Reduced Price Meals (FRPM) Count, total free meal count. Current Expense ADA (Average Daily Attendance) is the total expense of the average daily attendance (see Appendix A for more details).

Variable	Total daily energy	Operational	Non-operational
Area (sf)	100%	100%	100%
Enrollment	100%	100%	100%
Weekend	100%	100%	8%
School type	100%	100%	100%
Charter School	100%	92%	100%
High grade	100%	83%	100%
Current expense ADA	67%	92%	42%
Standard pressure (Hg)	67%	58%	58%
* Temperature	58%	58%	58%
** Free or Reduced priced meals	67%	58%	42%
Median household income	0%	0%	67%
Enrollment per area	17%	8%	83%

Warm weather temperature related variables were found to have very consistent impacts on each of the three metrics. Not only were they included in a total of seven months for each metric, but they were always included for the same months (April through October). These months span the warmest months throughout the state of California, thereby indicating that benchmarked scores for these months adjust to the increased temperatures experienced throughout the state. The other major weather related variable included was standard pressure. The selection of this variable is expected as it is a proxy for cloud cover given that low pressure weather systems are associated with clouds and potentially precipitation, resulting in lower levels of sunshine and solar irradiation.

The influence plots for the weather-related variables for the summer months for total daily energy consumption reveal their significant impact on school building energy use. For example, Fig. 2 illustrates the influence weather-related variables have across the building sample population for the month of September 2014. September was selected for illustrative purposes but we note that both temperature and pressure had similar effects throughout all months they were selected (April through October for temperature). The squared term for *cooling degree days* (CDD_2) is shown to be much more impactful for low performing buildings compared to efficient buildings. This indicates that these poor performers are highly sensitive to hot weather and may need a new AC unit or enhanced insulation to reduce this sensitivity. Our proposed DMT framework would enable these poor performing buildings to deepen their analysis and explore the daily fluctuations in their scores, like those between operational and non-operational states, to further unpack the source of their inefficiencies and develop mitigation strategies. Additionally, in Fig. 2 the *standard pressure* is shown to have the inverse influence on the building population with an increasingly negative affect for poor performing buildings. In other words, as pressure falls and corresponding cloud cover increases, lower performing buildings' energy usage increases at a much higher rate than their higher performing counterparts. We postulate that this is due to fact that lower performing buildings are currently served by inefficient indoor lighting systems that are utilized at high rates during cloudy days. This also means that low performing buildings could save significantly more energy when their lights are turned off, compared to efficient buildings, and could benefit more from new and efficient lighting systems.

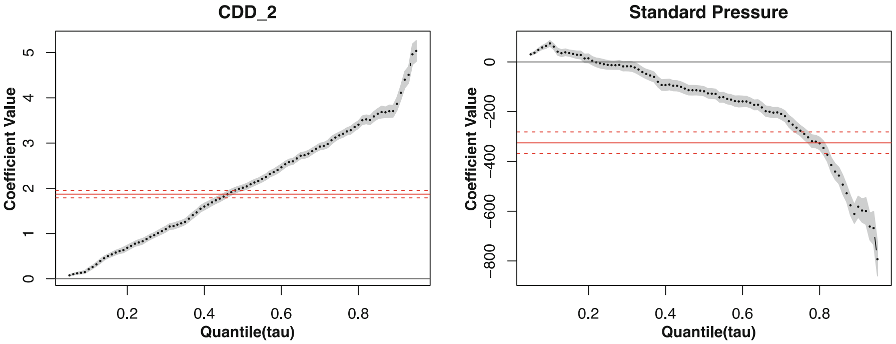


Fig. 2. Influence plots for weather related variables for September 2014. CDD_2 is the cooling degree days squared while Standard Pressure is measured in Hg.

We further deepen our analysis of energy drivers in our building sample population by exploring other influence plots (Fig. 3) for the most impactful variables with respect to total daily energy consumption. Figure 3 specifically provides plots for November 2014 to illustrate the effects of each variable but we note that the plots for other months display similar characteristics. Our results indicate that both *enrollment* and *area* (ft^2) are selected and show positive impacts on total daily energy use. However, the slope of their curves differ from each other. *Enrollment* is seen to have a positive impact throughout the building population (as seen by the positive coefficients) meaning that adding students to a school has significant impact for both well and poor performing buildings. This trend is likely due the fact that well performing buildings have been highly optimized to their population size and adding additional enrollment would push their energy usage upwards (albeit not vey much) and poor performing buildings have inherently inefficient systems that would require significant energy input to accommodate a growing occupant population. *Area* exhibits the biggest impact occurring on inefficient buildings and only a minimal impact on efficient buildings. This trend indicates that the size of a school building is not a strong indicator of energy performance for well performing buildings as the efficiencies of these major building systems can be agnostic to size.

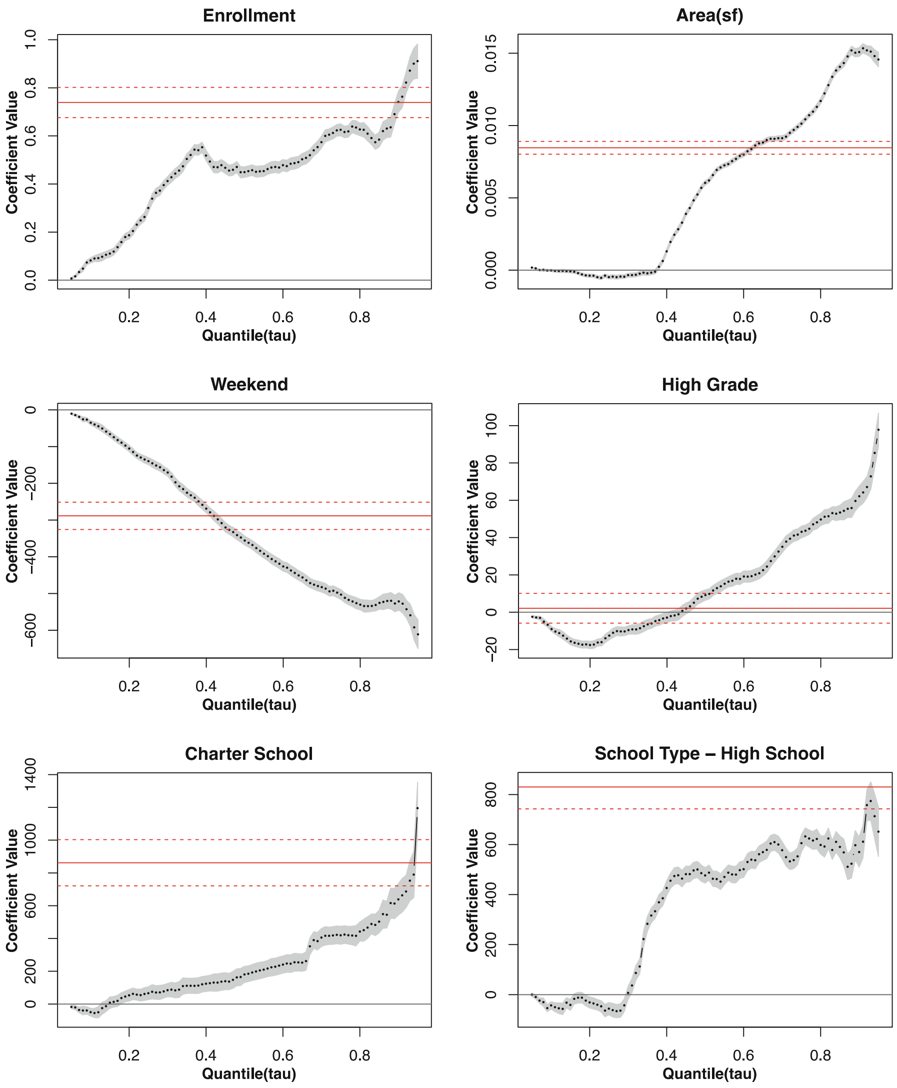


Fig. 3. Influence plots for most impactful variables for November 2014

Additionally, *weekend days* have a strong negative impact for building energy use, especially for poor performers. This impact may be driven by a larger decrease in energy use during weekends when building systems are functioning at low levels. For those high performing buildings, their efficient building systems may not cause the same drop in energy consumption between weekdays and weekends. *Charter schools* are also seen to be large users of energy, which may result from the fact that many of these schools lease their buildings rather than own them like other traditional schools in California. As a result, the incentive to implement energy efficiency retrofits is largely lost when the tenant does not own the building. *High school buildings* have a large impact on energy use when compared to elementary schools as shown by the coefficients for both the high grade and school type variables. Typically, high schools possess many more appliances that are needed to meet the needs of students in various science and computer based classes. Moreover, the inclusion of the school type variable further allows for comparison across all school types as scores for high school buildings are adjusted accordingly.

Examining the influence plots for the operational state reveals some additional insights into the energy dynamics of the school buildings. Figure 4. illustrates the effects of two variables, *current expense ADA* (Average Daily Attendance) and *FRPM* (Free or Reduced Price Meals) *count*, which are two proxies for school wealth. Current expense ADA has a positive impact on energy use, especially for those lower performing schools. Schools that have more money to spend on students may be purchasing more expensive computer equipment or have more sophisticated systems to condition the building. Moreover, schools that need to financially support kids' lunches through free or reduced meal plans (FRMP count) are also shown to use less energy.

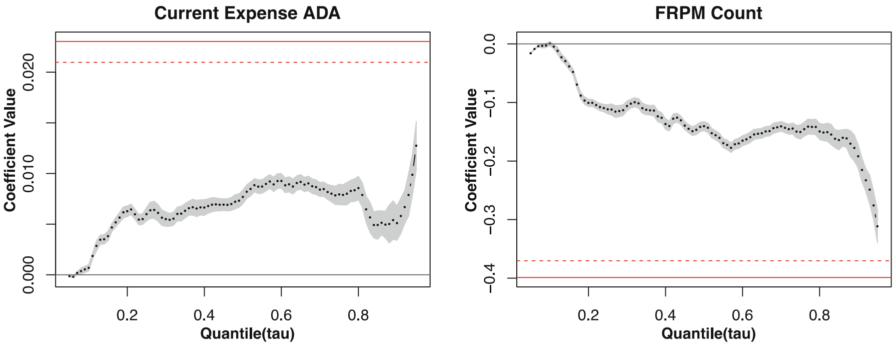


Fig. 4. Influence plots for current expense ADA and FRPM Count for the operational state for November 2014.

In order to further demonstrate the merits of our proposed DMT benchmarking framework, we explore how the differences in daily performance scores over time can drive additional insights. Figure 5. depicts the total daily energy use scores for one sample school building (located in Fresno, CA) across an entire year. The weekend scores (blue-filled squares) are much more consistent than the weekday scores (red-open diamonds), suggesting that the building is not used during the weekends (adjusting to a baseline load) and that the building systems properly adapt to the low occupancy. The lower weekday scores suggest that the building systems may be relatively inefficient compared to the other school buildings in the dataset. The lower (and worse) scores in the summer months supports this notion as the building becomes more inefficient during hot days, implying a potential inefficient HVAC unit.

Figure 6. depicts the scores for both the operational (red-open diamonds) and non-operational (blue-filled squares) scores for the same high school in Fresno, California

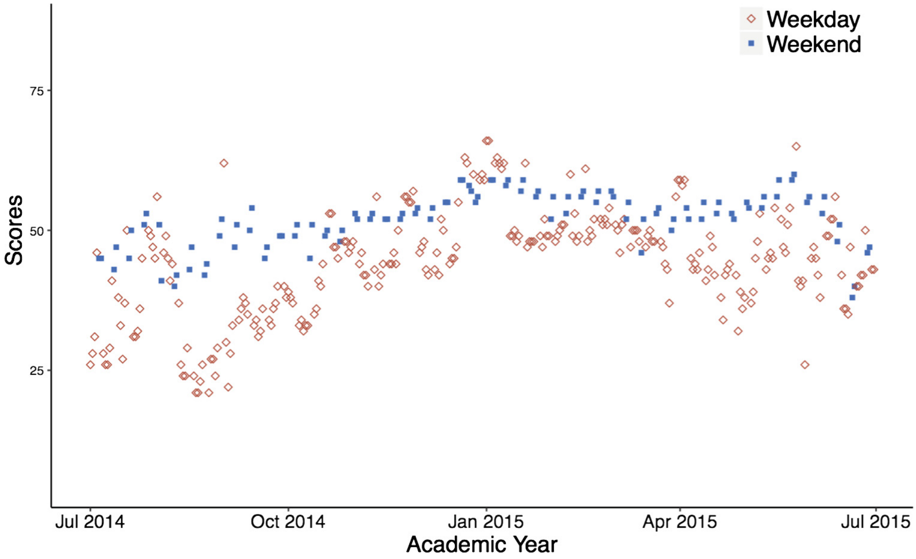


Fig. 5. Total daily energy scores for a high school in Fresno, California. Red points indicate weekday scores while blue points represent weekend scores. (Color figure online)

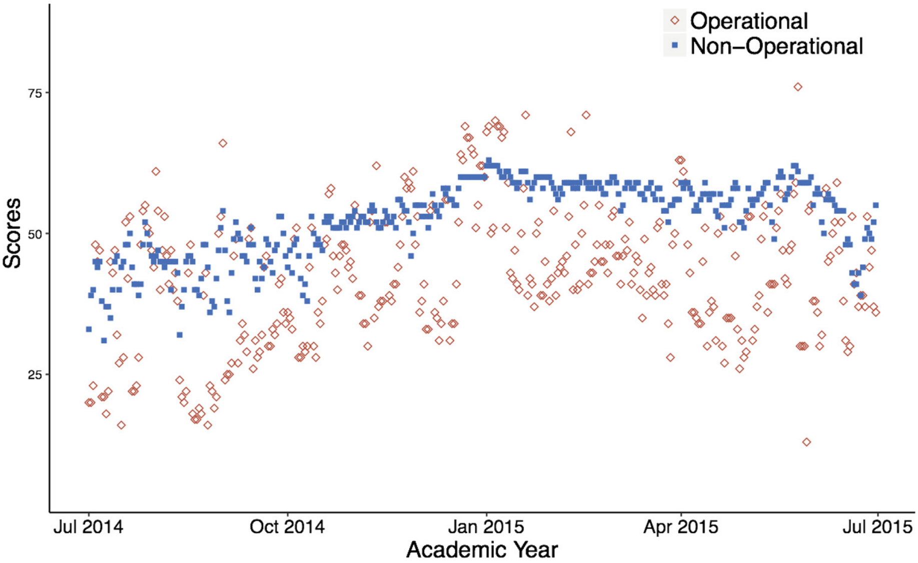


Fig. 6. Operational and Non-Operational scores for the same high school shown in Fig. 5. Red points indicate Operational scores while blue points indicate Non-Operational scores. (Color figure online)

that is shown in Fig. 5. The non-operational state scores follow a similar trend to the weekend scores from Fig. 5., ranging from the mid-50s during the fall months and rising to the low 70s during the winter months. The consistent scores between both metrics supports the previous theory that the building control systems are functioning properly. In contrast, a school that shows poor weekend scores but good weekday scores may have control system issues. Both figures also show a rise in operational and weekday scores for the last few days of December and first few days of January. These days correspond to winter break when the school building is largely unoccupied, which explains the increased performance for those days; several of the other low operational state scores also correspond to school break days.

6 Limitations and Future Work

While this paper aims to propose a novel DMT energy benchmarking framework that overcomes several challenges faced by previous work, several limitations persist. As with all energy benchmarking models, verification and validation is difficult as it is infeasible to manually audit and check all buildings in our sample dataset in order to establish their inefficiencies. Future work aims to take a first-step towards addressing this issue by validating the scores derived from this model by conducting interviews with select building managers at different schools throughout the state of California and comparing our model's results against the qualitative results of these interviews. Currently, we are in the process of contacting managers at schools of interest and are planning to disseminate a survey to reach a broader audience.

Additionally, the multitude of scores produced by this model provides normalized metrics that allows for comparison of vastly different school buildings in different climates. However, the presented analysis is limited to a specific building type (i.e., school buildings) with similar and known operational schedules. The structure of this framework can be more broadly applied and does allow for the inclusion of additional features (e.g., building type); the model ascribes interpretable effects to each feature, but they may require careful construction to adequately account for differences in uses and operational schedules between building types which are not captured by the other covariates. As currently presented, the scores from the DMT framework are also limited in their applicability to fault detection and diagnosis (FDD) in buildings. The DMT scores are aimed to be a starting point for the development of future fault detection and diagnosis algorithms concerning building-level operations using publicly available data. The relative scores of these metrics for buildings can be compared to identify differences in performance during these states and can be supplemented with raw energy consumption values to aid in building-level fault detection methods. Future work aims to extend the benchmarked metrics to include peak power consumption, as this value is often used in utility billing and can make up a substantial fraction of the bill. Furthermore, understanding the peak power consumption of buildings on a daily scale can help building managers isolate and shift system operations to reduce this value.

7 Conclusion

The proposed Data-driven, Multi-metric, and Time-varying (DMT) framework takes advantage of emerging smart meter infrastructure and new public data streams (building characteristics, social factors) to analyze a large portfolio of buildings at the daily scale and across several metrics. The DMT framework extends previous work by using quantile regression to model the entire conditional distribution and downplay the impact of outlier data. This enables deeper insights into low and high performing buildings by differentiating the impact between their energy drivers. Daily energy benchmarking further allows for the use of daily weather data, thereby extracting more information from readily available data sources and allowing for seasonally adjusted scores. In practice, daily scores can provide building managers and utilities with more immediate and reliable feedback on performance, facilitating in faster detection of abnormal building operations and increased educational effects. Beyond more frequent benchmarking, the DMT framework examines operational and non-operational performance, enabling the comparison of intra-day energy consumption that can provide deeper insights into building operations. Due to the large human component of building design and operations, the DMT framework was designed to maintain interpretability so that utilities, municipalities, and building managers alike could gain deeper insights into building performance to empower them to make more informed decisions.

To illustrate the types of insights that the DMT framework can provide, it was tested on an integrated dataset consisting of smart meter energy, daily weather, and building characteristics of schools throughout the state of California. Results from this study showed that there were consistent drivers of energy consumption throughout the year for school buildings; these drivers included area, enrollment, school type, charter schools, and high grade. Additionally, temperature related variables were deemed to have a significant effect on energy consumption during warmer months. Examining the scores for buildings between metrics and over time displayed some added insights that would not be possible with conventional benchmarking models. Results from one school showed consistent performance during weekends and its non-operational state, indicating a well performing control system. However, worsening scores during summer months reveal that the building is sensitive to increases in temperature, suggesting the building has an inefficient HVAC system. Scores and influence plots for other buildings may exhibit different trends, demonstrating that individualized insights can be attained by using the DMT framework.

Historically, data limitations have restricted current energy benchmarking models to only examine total energy consumption, typically at the yearly scale. New widespread adoption of smart meters is, however, creating an opportunity to benchmark buildings with an information rich data stream on finer temporal scales. The proposed DMT energy benchmarking framework utilizes this new high-fidelity data stream to produce daily scores across several metrics, which can then be used to examine seasonal trends and provide insights that are undetected by current benchmarking models. Changes in building performance overtime can be studied more quickly with daily benchmarking, thereby enabling the construction of more sophisticated fault diagnosis and detection systems based on this data. An energy benchmarking framework that utilizes smart meter data can bridge the gap between data science and traditional engineering approaches by creating a more informative tool that empowers decision-makers.

Appendix A

Variable name	Characteristic and Explanation
District type	District Ownership Type Description
Educational type	Educational Option Type Description
Charter School	A “Y” or “N” value indicating whether a school is a charter school in the current academic year.
High grade	Highest grade offered
Enrollment	A total count of K-12 students enrolled (primary or short-term) on Census Day (the first Wednesday in October). These data were submitted to CALPADS as part of the annual Fall 1 submission
Total free meal count	Of the *Enrollment (K-12)*, a total unduplicated count of students who meet household income or categorical eligibility criteria for free meals based on one or more of the following reasons: (1) applying for the National School Lunch Program (NSLP); (2) submitting alternative household income forms; (3) student homeless or migrant statuses in CALPADS; (4) being “directly certified” through CALPADS as participating in California’s food stamp or CalWORKs programs during July - November; or (5) being identified through the weekly CALPADS Foster Matching process or matched by the LEA through the CALPADS online match process as being in Foster Placement or Foster Family Maintenance on Census day. The *Free Meal Count (K-12)* is not displayed on any CALPADS report; however, this count represents the official *Free Meal Count (K-12)* for the academic year.
Percent eligible free	The percent of students eligible for free meals. [*Free Meal Count (K-12)* divided by *Enrollment (K-12)*]
FRPM count	Of the *Enrollment (K-12)*, a total unduplicated count of students who meet household income or categorical eligibility criteria for free or reduced meals (FRPM) based on one or more of the following reasons: (1) applying for the National School Lunch Program (NSLP); (2) submitting alternative household income forms; (3) student homeless or migrant statuses in CALPADS; (4) being “directly certified” through CALPADS as participating in California’s food stamp or CalWORKs programs during July - November; or (5) being identified through the weekly CALPADS Foster Matching process or matched by the LEA through the CALPADS online match process as being in Foster Placement or Foster Family Maintenance on Census day

(*continued*)

(*continued*)

Variable name	Characteristic and Explanation
Percent eligble FRPM	The percent of students eligible for free or reduced price meals (FRPM). [*FRPM Count (K-12)* divided by *Enrollment (K-12)*]
EDP 365	The total cost for the current expense of education
Current expense ADA	Total ADA (average daily attendance) is defined as the total days of student attendance divided by the total days of instruction. This is the total cost of the ADA
Current expense per ADA	The total cost per ADA or the EDP_365 divided by the Current Expense ADA
School type	The type of school as either "High School", "Unified", or "Elementary"
Area (sf)	Total area of the school building(s) in square feet
Median Household Income	The median household income for the zip code taken from the US Census Bureau
Temperature max	The maximum temperature recorded during the day in Fahrenheit
Temperature min	The minimum temperature recorded during the day in Fahrenheit
Temperature mean	The average daily temperature for the day in Fahrenheit
Dewpoint	The average daily dewpoint temperature for the day in Fahrenheit
Temperature wetbulb	The average daily wetbulb temperature for the day
Heating degree day (HDD)	Number of degrees that the day's average temperature was below 65° Fahrenheit
Cooling degree day (CDD)	The number of degrees that the day's average temperature was above 65° Fahrenheit
Total precipitation	The total amount of precipitation (water equivalent) in inches
Standard pressure	The total standard pressure for the day in Hg
Result speed	The resulting wind speed for the day
Average wind speed	The daily average wind speed in miles per hour
Max5speed	The max speed of wind with a duration of 5 min
Max2speed	The max speed of wind with a duration of 2 min
Temperature mean squared	The average daily temperature squared
Heating degree day squared (HDD_2)	The heating degree day (HDD) squared
Cooling degree day squared (CDD_2)	The cooling degree day (CDD) squared
Temperature mean natural log	The natural log transformation of the average daily temperature
Weekend	Dummy variable where "1" is a weekend and "0" is a weekday
Enrollment per area	The total enrollment per unit area (Students per square foot)

References

1. Nowak, S., Baatz, B., Gilleo, A., Kushler, M., Molina, M., York, D.: Beyond carrots for utilities: a national review of performance incentives for energy efficiency, Washington, DC (2015). http://kms.energyefficiencycentre.org/sites/default/files/u1504.pdf. Accessed 31 May 31 2017
2. U.S. Department of Energy, Buildings energy databook, Silver Spring (2012). http://buildingsdatabook.eren.doe.gov/DataBooks.aspx. Accessed 24 Sept 2017
3. Ramesh, T., Prakash, R., Shukla, K.K.: Life cycle energy analysis of buildings: an overview. Energy Build. **42**, 1592–1600 (2010). https://doi.org/10.1016/j.enbuild.2010.05.007
4. National Academy of Sciences, Real Prospects for Energy Efficiency in the United States: America's Energy Future Panel on Energy Efficiency Technologies, National Academies Press, Washington, D.C. (2010). https://doi.org/10.17226/12621
5. How many smart meters are installed in the United States, and who has them? U.S. Energy Information Administration (EIA) (2017). https://www.eia.gov/tools/faqs/faq.php?id=108&t=3. Accessed 10 Jan 2018
6. Map: U.S. Building Benchmarking and Transparency Policies, Institute for Market Transformation (2017). http://www.imt.org/resources/detail/map-u.s.-building-benchmarking-policies. Accessed 10 May 2017
7. Lutzenhiser, L., Moezzi, M., Hungerford, D., Commission, C.E., Friedmann, R., Gas, P., Company, E.: Sticky points in modeling household energy consumption defining the problem. In: 2010 ACEEE Summer Study on Energy Efficiency in Buildings, pp. 167–182. (2010) https://www.pdx.edu/cus/sites/www.pdx.edu.cus/files/Lutzenhiseret al (2010) Sticky Points in Modeling Household Energy Consumption.pdf. Accessed 14 Sept 14 2017
8. Russell, R., Guerry, A.D., Balvanera, P., Gould, R.K., Basurto, X., Chan, K.M.A., Klain, S., Levine, J., Tam, J.: Humans and nature: how knowing and experiencing nature affect well-being. Annu. Rev. Environ. Resour. **38**, 473–502 (2013). https://doi.org/10.1146/annurev-environ-012312-110838
9. Karlin, B., Zinger, J.F., Ford, R.: The effects of feedback on energy conservation: a meta-analysis. Psychol. Bull. **141**, 1205–1227 (2015). https://doi.org/10.1037/a0039650
10. Jain, R.K., Taylor, J.E., Culligan, P.J.: Investigating the impact eco-feedback information representation has on building occupant energy consumption behavior and savings. Energy Build. **64**, 408–414 (2013). https://doi.org/10.1016/j.enbuild.2013.05.011
11. Filippín, C.: Benchmarking the energy efficiency and greenhouse gases emissions of school buildings in central Argentina. Build. Environ. **35**, 407–414 (2000). https://doi.org/10.1016/S0360-1323(99)00035-9
12. Chung, W.: Review of building energy-use performance benchmarking methodologies. Appl. Energy **88**, 1470–1479 (2011). https://doi.org/10.1016/j.apenergy.2010.11.022
13. Xuchao, W., Priyadarsini, R., SiewEang, L.: Benchmarking energy use and greenhouse gas emissions in Singapore's hotel industry. Energy Policy **38**, 4520–4527 (2010). https://doi.org/10.1016/j.enpol.2010.04.006
14. U.S.E.P. Agency, ENERGY STAR score technical reference, pp. 1–14 (2014). https://portfoliomanager.energystar.gov/pdf/reference/ENERGY STAR Score.pdf. Accessed 7 Sept 2017
15. Farrell, M.J.: The measurement of productive efficiency. J. R. Stat. Soc. Ser. A (General) **120**, 253–290 (1957). https://doi.org/10.1016/S0377-2217(01)00022-4
16. Zhou, P., Ang, B.W., Poh, K.L.: A survey of data envelopment analysis in energy and environmental studies. Eur. J. Oper. Res. **189**, 1–18 (2008). https://doi.org/10.1016/j.ejor.2007.04.042

17. Schmidt, P.: Frontier production functions. Econom. Rev. **4**, 289–328 (1985). https://doi.org/10.1080/07474938608800089
18. Lee, W.L., Burnett, J.: Benchmarking energy use assessment of HK-BEAM, BREEAM and LEED. Buil. Environ. **43**, 1882–1891 (2008). https://doi.org/10.1016/j.buildenv.2007.11.007
19. Kavousian, A., Rajagopal, R.: Data-driven benchmarking of building energy efficiency utilizing statistical frontier models. J. Comput. Civil Eng. **28**, 79–88 (2014). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000327
20. Filippini, M., Hunt, L.C.: US residential energy demand and energy efficiency: a stochastic demand frontier approach. Energy Econ. **34**, 1484–1491 (2012). https://doi.org/10.1016/j.eneco.2012.06.013
21. Buck, J., Young, D.: The potential for energy efficiency gains in the Canadian commercial building sector: a stochastic frontier study. Energy **32**, 1769–1780 (2007). https://doi.org/10.1016/j.energy.2006.11.008
22. Yang, Z., Roth, J., Jain, R.K.: DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy Build. **163**, 58–69 (2018). https://doi.org/10.1016/J.ENBUILD.2017.12.040
23. Arora, D., Malik, P.: Analytics: key to go from generating big data to deriving business value. In: 2015 IEEE First International Conference on Big Data Computing Service and Applications, pp. 446–452. IEEE (2015). https://doi.org/10.1109/bigdataservice.2015.62
24. Kavousian, A., Rajagopal, R., Fischer, M.: Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy **55**, 184–194 (2013). https://doi.org/10.1016/j.energy.2013.03.086
25. Haben, S., Singleton, C., Grindrod, P.: Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Trans. Smart Grid **7**, 136–144 (2016). https://doi.org/10.1109/TSG.2015.2409786
26. Kwac, J., Flora, J., Rajagopal, R.: Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid **5**, 420–430 (2014). https://doi.org/10.1109/TSG.2013.2278477
27. Alahakoon, D., Yu, X.: Smart electricity meter data intelligence for future energy systems: a survey. IEEE Trans. Industr. Inf. **12**, 425–436 (2016). https://doi.org/10.1109/TII.2015.2414355
28. Chicco, G.: Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy **42**, 68–80 (2012). https://doi.org/10.1016/j.energy.2011.12.031
29. Zhao, F., Wang, G., Deng, C., Zhao, Y.: A real-time intelligent abnormity diagnosis platform in electric power system. In: International Conference on Advanced Communication Technology, ICACT, Global IT Research Institute (GIRI), pp. 83–87 (2014). https://doi.org/10.1109/icact.2014.6778926
30. Albert, A., Rajagopal, R.: Smart meter driven segmentation: what your consumption says about you. IEEE Trans. Power Syst. **28**, 4019–4030 (2013). https://doi.org/10.1109/TPWRS.2013.2266122
31. Beckel, C., Sadamori, L., Staake, T., Santini, S.: Revealing household characteristics from smart meter data. Energy **78**, 397–410 (2014). https://doi.org/10.1016/j.energy.2014.10.025
32. Dent, I., Aickelin, U., Rodden, T., Craig, T.: Finding the creatures of habit; clustering households based on their flexibility in using electricity. SSRN Electron. J. (2012). https://doi.org/10.2139/ssrn.2828585

33. Cao, H.A., Beckel, C., Staake, T.: Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns. In: IECON Proceedings (Industrial Electronics Conference), pp. 4733–4738. IEEE (2013). https://doi.org/10.1109/iecon.2013.6699900
34. Magoulès, F., Zhao, H., Elizondo, D.: Development of an RDP neural network for building energy consumption fault detection and diagnosis. Energy Build. **62**, 133–138 (2013). https://doi.org/10.1016/j.enbuild.2013.02.050
35. Yu, Y., Woradechjumroen, D., Yu, D.: A review of fault detection and diagnosis methodologies on air-handling units. Energy Build. **82**, 550–562 (2014). https://doi.org/10.1016/j.enbuild.2014.06.042
36. Li, S., Wen, J.: A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. Energy Build. **68**, 63–71 (2014). https://doi.org/10.1016/j.enbuild.2013.08.044
37. Liang, J., Du, R.: Model-based fault detection and diagnosis of HVAC systems using support vector machine method. Int. J. Refrig. **30**, 1104–1114 (2007). https://doi.org/10.1016/j.ijrefrig.2006.12.012
38. Du, Z., Fan, B., Jin, X., Chi, J.: Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Build. Environ. **73**, 1–11 (2014). https://doi.org/10.1016/j.buildenv.2013.11.021
39. Capozzoli, A., Lauro, F., Khan, I.: Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst. Appl. **42**, 4324–4338 (2015). https://doi.org/10.1016/j.eswa.2015.01.010
40. Bynum, J.D., Claridge, D.E., Curtin, J.M.: Development and testing of an automated building commissioning analysis tool (ABCAT). Energy Build. **55**, 607–617 (2012). https://doi.org/10.1016/j.enbuild.2012.08.038
41. Wang, H., Xu, P., Lu, X., Yuan, D.: Methodology of comprehensive building energy performance diagnosis for large commercial buildings at multiple levels. Appl. Energy **169**, 14–27 (2016). https://doi.org/10.1016/j.apenergy.2016.01.054
42. Grueneich, D., Jacot, D.: Scale, Speed, and persistence in an analytics age of efficiency: how deep data meets big savings to deliver comprehensive efficiency. Electr. J. **27**, 77–86 (2014). https://doi.org/10.1016/j.tej.2014.03.001
43. Larsen, P.H., Carvallo, J.P., Goldman, C.A., Murphy, S., Stuart, E., Rockwell, K., Schell, S., Nicholls, L.: Updated Estimates of the Remaining Market Potential of the U.S. ESCO Industry (2017). https://emp.lbl.gov/sites/default/files/revised_market_potential_final_25apr2017.pdf. Accessed 8 May 2017
44. International Performance Measurement and Verification Protocol Concepts and Options for Determining Energy and Water Savings, Toronto, CA, vol. 1 (2012). http://www.eeperformance.org/uploads/8/6/5/0/8650231/ipmvp_volume_i__2012.pdf. Accessed 1 Oct 2017
45. FEMP, M&V Guidelines: Measurement and Verification for Performance-Based Contracts - Version 4.0 (2015). http://www1.eere.energy.gov/. Accessed 27 Dec 2017
46. Ke, M.T., Yeh, C.H., Jian, J.T.: Analysis of building energy consumption parameters and energy savings measurement and verification by applying eQUEST software. Energy Build. **61**, 100–107 (2013). https://doi.org/10.1016/j.enbuild.2013.02.012
47. Granderson, J., Touzani, S., Fernandes, S., Taylor, C.: Application of automated measurement and verification to utility energy efficiency program data. Energy Build. **142**, 191–199 (2017). https://doi.org/10.1016/j.enbuild.2017.02.040
48. Heo, Y., Zavala, V.M.: Gaussian process modeling for measurement and verification of building energy savings. Energy Build. **53**, 7–18 (2012). https://doi.org/10.1016/j.enbuild.2012.06.024

49. Burkhart, M.C., Heo, Y., Zavala, V.M.: Measurement and verification of building systems under uncertain data: a Gaussian process modeling approach. Energy Build. **75**, 189–198 (2014). https://doi.org/10.1016/j.enbuild.2014.01.048
50. Roth, J., Rajagopal, R.: Benchmarking building energy efficiency using quantile regression. Energy (2018). https://doi.org/10.1016/j.energy.2018.02.108
51. Koenker, R., Bassett, G.: Regression quantile (1978). https://doi.org/10.1257/jep.15.4.143
52. Chen, C.: An introduction to quantile regression and the QUANTREG Procedure. Sugi **30**, 1–24 (2001)
53. White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica **48**, 817–838 (1980). https://doi.org/10.2307/1912934
54. Ürge-Vorsatz, D., Cabeza, L.F., Serrano, S., Barreneche, C., Petrichenko, K.: Heating and cooling energy trends and drivers in buildings. Renew. Sustain. Energy Rev. **41**, 85–98 (2015). https://doi.org/10.1016/j.rser.2014.08.039
55. Landerman, L.R., Land, K.C., Pieper, C.F.: An empirical evaluation of the predictive mean matching method for imputing missing values. Sociol. Methods Res. **26**, 3–33 (1997). https://doi.org/10.1177/0049124197026001001

Visual Pattern Recognition as a Means to Optimising Building Performance?

Tristan Gerrish[1], Kirti Ruikar[2](✉), Malcolm Cook[2], Mark Johnson[1], and Mark Philip[1]

[1] BuroHappold, 230 Lower Bristol Road, Bath BA2 3DQ, UK
[2] Loughborough University, Lougborough LE113TU, UK
k.d.ruikar@lboro.ac.uk

Abstract. The move towards the use of smart systems to record data about the performance of building presents an opportunity to examine patterns of behaviour, which shed light on its performance peaks and troughs. Answers often lie within these peaks and troughs. For example, smart systems are used to record in-use data such as room temperatures, thermal comfort, and lighting usage. Each system is designed for an expected ‘behaviour pattern’ that is bounded by a ‘threshold’ boundary. This is the expected performance the system is designed for. Any deviations in this performance may indicate of a system malfunction, its overuse or underuse due to unexpected usage of building space/s (e.g. large number of visitors, doors/windows being left open). Performance when bounded within the threshold limits would be considered to be ‘normal’ or ‘expected’. Thus, answers fall within four categories of performance behaviours. These include system malfunction, system overuse, system underuse, and normal performance. As data and information are accumulated over a period of time they present opportunities to observe system behaviour patterns and present opportunities to map these patterns and classify within learning clusters of ‘expected’ and ‘unexpected’ thresholds. Doing so would enable building owners to truly understand the building, so performance can be firstly understood and then optimised. Thus, an unusual activity that is significantly beyond the expected ‘norm’ would present an opportunity to learn about the building so a healthy function can be determined and maintained. The generation of large datasets through extensive monitoring has created a potential environment in which big-data style analytics could be applied for holistic performance assessment and pattern recognition. This research builds on the work previously completed by Gerrish et al. [1–4, 21] and utilises techniques of visualisation to demonstrate such behaviour patterns and presents learning opportunities for optimal performance. This is demonstrated through visualisation of energy performance data for a case-study building in the UK.

Keywords: Energy performance management · Visualisation · BIM facilities Management

© Springer International Publishing AG, part of Springer Nature 2018
I. F. C. Smith and B. Domer (Eds.): EG-ICE 2018, LNCS 10863, pp. 594–618, 2018.
https://doi.org/10.1007/978-3-319-91635-4_31

1 Introduction

The move towards the use of smart systems to record data about the performance of building presents an opportunity to examine patterns of behaviour, which shed light on its performance peaks and troughs. Answers often lie within these peaks and troughs. For example, smart systems are used to record in-use data such as room temperatures, thermal comfort, and lighting usage. Each system is designed for an expected 'behaviour pattern' that is bounded by a 'threshold' boundary. This is the expected performance the system is designed for. Any deviations in this performance may indicate of a system malfunction, its overuse or underuse due to unexpected usage of building space/s (e.g. large number of visitors, doors/windows being left open). Performance when bounded within the threshold limits would be considered to be 'normal' or 'expected'. Thus, answers fall within four categories of performance behaviours. These include system malfunction, system overuse, system underuse, and normal performance. As data and information are accumulated over a period of time they present opportunities to observe system behaviour patterns and present opportunities to map these patterns and classify within learning clusters of 'expected' and 'unexpected' thresholds. Doing so would enable building owners to truly understand the building, so performance can be firstly understood and then optimised. Thus, an unusual activity that is significantly beyond the expected 'norm' would present an opportunity to learn about the building so a healthy function can be determined and maintained. The generation of large datasets through extensive monitoring has created a potential environment in which big-data style analytics could be applied for holistic performance assessment and pattern recognition. This research builds on the work previously completed by Gerrish et al. [1–4] and utilises techniques of visualisation to demonstrate such behaviour patterns and presents learning opportunities for optimal performance. This is demonstrated through visualisation of energy performance data for a case-study building in the UK.

2 Case-Study

This research explored how BIM could be used to support non-domestic building Energy Performance Management (EPM). For this purpose, an existing building was used as a case-study for research development and application. This building was fitted with a comprehensive Building Management System (BMS) to control systems and record energy consumption data (e.g. CO_2, temperature, lighting, etc.). The original design documentation and models of the building were available from the onset; however, since the original design was later subject to changes and updated versions were no longer available, it resulted in inaccuracies between the available models and the as-built building. To avoid misinterpretations as a result of inaccuracies, new 'representative' models of the as-built building were created. This was thought necessary because by modelling a building's form, function and constituent systems, the evaluation of design decisions and their impact on a completed building can be better understood and contextualised. Measuring how each of these factors contributes toward its performance provides the designer with an improved understanding of how their decisions impact on the conditional aspects of performance. Thus, the main tasks performed at this stage were, (1) update and consolidate building energy performance related models and information generated during design, using information gathered from the case-study project's design development directory; and (2) create an as-built modelled representation of the in-use building. This work was undertaken to provide a basis for further research into the operational performance management and investigation of a case-study building using a performance data attributed BIM.

3 The Case-Study Building

The case-study building (see Fig. 1) is a high-performance office located in the North of England. BuroHappold Engineering contributed to the building's mechanical, structural, ground, fire and façade engineering designs in addition to providing specialist consulting on security, lighting design, acoustics, sustainability and environment. The primary reason for use of this building was the extent to which it was to be monitored upon completion, from which operational performance information could be obtained. The building is comprised of open-plan offices around a large atrium, conditioned using a complex Heating, Ventilation and Air-Conditioning (HVAC) system fed from two Combined Heat and Power (CHP) engines with gas boiler backups, propane and absorption chillers in conjunction with passive cooling towers and ground tempered ventilation. Space heating and cooling is achieved via perimeter trench heating and passive chilled beams across all floors above ground level, with Fan Coil Units (FCUs) supplementing this at ground floor and in server rooms. These systems are controlled via a BMS with 28 distinct modes of operation depending on external conditions and demand for heating and cooling.

At research commencement, the building was in the process of final fit-out and commissioning, prior to handover and occupation in March 2013. Building geometry information was available in Revit format for architectural and structural detailing; however, mechanical services and building energy performance data were only available in a 2D CAD and the proprietary IES-VE (building energy performance modelling and simulation program) format, respectively. Engineering data and level of detail for each discipline (constituting a description of holistic building performance) available at research commencement is detailed in Table 1. Project files containing all building models, building specifications and other supporting documentation and details about the building's composition, as-designed performance criteria and the methods of operation were obtained and analysed for the purpose of the case-study.

BuroHappold's role within the building's design process was that of a consultant engineer providing services during RIBA [5] Stages 1–4 (with input into Stage 7 for commissioning and sign-off). Creation of system accurate as-built models during these stages were not a priority and for which creation would not be feasible given the likelihood for changes in construction and handover. Research into the use of BIM for managing building energy performance required the building to be supported by information stored in a BIM environment. However, the building was completed at the commencement of the BIM adoption strategy and models other than partial architectural and superseded structural models were unavailable. The client did not specify a

Fig. 1. Case-study Building [50]

BIM delivered project and as such the majority of the information describing the case-study building was primarily in the form of spreadsheets and drawings. The information gathered from handover documentation and from design and commissioning teams informed the development of further ‘enriched’ models in IES-VE and Autodesk Revit to a standard suitable for use in later research.

4 Adopted Approach

4.1 Document Review

Document review was the primary method of collecting information describing the case-study building. Bowen [6] explains the justification for documentation review as an inexpensive and unobtrusive method of gaining background information that provides a ‘behind the scenes’ view of information available through more prominent sources such as end of work stage reports and/or most recent models. However, a thorough review of all information generated during design could be too time consuming given that it can be subject to bias from selective information survival [7] (where relevant information is not found due emphasis on the documentation of more favourable information, the potential for incomplete or inaccurate records [8] and often disorganized [9–11].

Initially, the models developed by the design teams and their supporting documentation were obtained. These provided a comprehensive dataset detailing the building’s composition and its intended energy performance. A meta-analysis of the

Table 1. Building models available at research commencement

Discipline	Modelling environment[a]	Modelled extent

Architectural design	Autodesk Revit Architecture	Geometry (based on preliminary Sketchup models);
		Orientation;
		Fabric specification;
		Junction detailing;
		Space scheduling; and
		Materials scheduling (not including performance specification)

Structural engineering	Autodesk Revit Structures	Geometry (based on Architectural specification); and Element performance specifications (from Autodesk Robot and Tekla Structures)

MEP engineering	AutoCAD	Geometry (for detailed junctions only);
		Electrical systems layout;
		Ventilation systems layout;
		CHW layout (including HTCHW and LTCHW);
		DHWS layout; and
		All system performance specifications derived from discipline specific tools and calculations to Stage D detail

EPM and sustainability engineering	IES-VE	Location;
		Geometry (simplified for EPM constraints);
		Fabric performance specifications;
		Heating system characteristics;
		Cooling system characteristics
		Ventilation system characteristics; and
		Space-type based thermal profiles

[a] Modelling environment in this context is the principal data storage mechanism in which information describing a building's composition and systems is recorded. Modelling environment may also refer to platforms in which this data is generated, such as the IES-VE package and specific MEP design tools; however, these mainly serve to generate such information and provide reasoning for its implementation.

project's documentation structure was undertaken showing the extent of information generated throughout design (Fig. 2). The majority of this information would likely be out-of-date upon building handover due to changes made throughout the design development, and the length of time between creation and utilisation of such information. Large numbers of simulation files created at early stages are likely disproportionate due to the number of documents created to support analysis between Preparation and Developed Design.

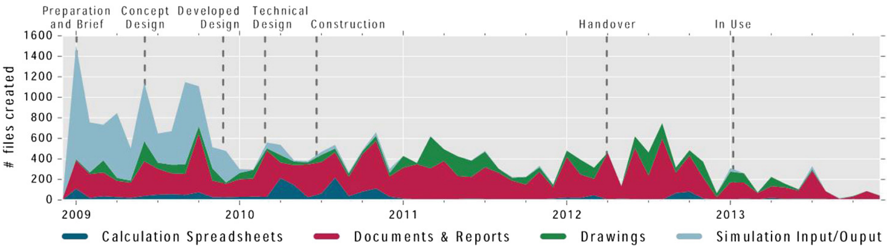


Fig. 2. Design documentation type production frequency

4.2 Preparation and Developed Design

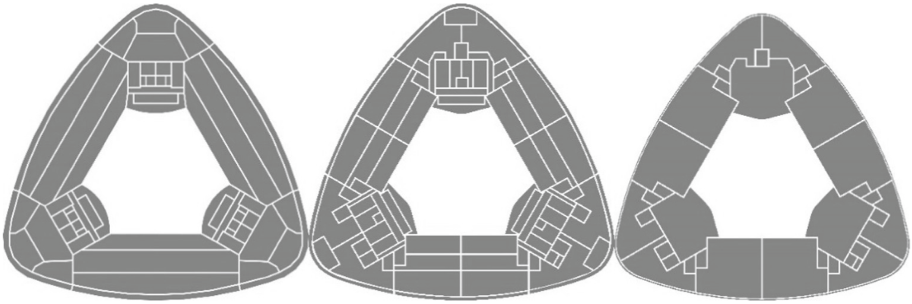
Reports signifying handover of information for the next stage of design provided the primary source of information describing the building, and expected levels of performance; however, following handover to the specialist contractors at the Technical Design stage, changes became less documented. Handover documentation and drawings compiled into Operations and Maintenance (O&M) manuals provided the most concise and structured source of information from which to draw accurate designed performance and configuration data. However, just as Gallaher [12] found, this too contained errors from corruption in storage media and out-of-date documentation. Reliance on existing data would have likely produced further error later in research, providing justification for manual recreation of building performance and BIM environments.

4.3 Baseline Performance Model

The most up to date building energy performance model that was available for this research was completed 2 years prior to construction. There were several differences between the ‘as-designed’ and ‘as-built’ building, which would most likely result in gaps between predicted and in-use performance if comparisons were to be made.

Differences Between Design and In-Use Performance Depiction
A substantial difference was identified between how prediction and measurement of the building’s performance differed. For example, the partial layout used in simulation simplified geometry to reduce simulation and modelling times; however, operational monitoring divided spaces in a completely different way for sub-metering (Fig. 3). These differences demonstrate one of the challenges in linking BIM and EPM, correlating their often conflicting modelling requirements across discipline domains [13, 14].

Creating an As-Built Performance Model
Visual representation of complex operation of spatial conditioning in an energy performance model was attempted, incorporating changes found in the operations of the building. Using IES-VE [15], a space-wise simulation of air supply and extract in conjunction with calibration of space-based electrical loading using available BMS monitored performance information was attempted; however, the complexity of the case-study building’s HVAC systems prevented completely accurate modelling of



(a) Simplified EPM layout for simulation efficiency (61 spaces)

(b) EPM layout required for accurate simulation of spatial performance (79 spaces)

(c) Actual building control layout used by the BMS (36 spaces)

Fig. 3. Spatial delineation differences for simulation and operation of a building's systems on a representative floor-plan

these and would have required excessive time in model calibration [13], for which there was little justification in the context of this research. A simplified representation of the completed building's composition was created without definition of the specific systems providing heating, cooling and ventilation, resulting in a less detailed and more general prediction of whole buildings energy performance, for the purpose of providing a baseline for comparison. The method used for creation of the representative model is defined by the U.S. Department of Energy [16] which gives guidance for evaluation of holistic building energy. In particular, this method was chosen given the substantive effects of each complex HVAC system of operation, making isolation of these effects complex beyond the purpose of providing a baseline energy performance model.

4.4 BIM Environment for Performance Data Interaction

This research began (in 2014) at the time when the mandate for BIM with its 2016 target for Level 2 BIM adoption had prompted many UK companies to take measures towards BIM readiness. This period also saw a drive from the Government to achieve energy efficiency targets, which prompted several buildings being designed to BREEAM standards. The case-study building was a BREEAM outstanding building, which presented a 'test-bed' opportunity to examine, if/how data from energy simulation models could be linked to a BIM. The thinking at the time being that if the target of the industry had been to ultimately achieve 'single model capability', then this would be an effort to unify the currently disparate datasets, so process efficiency could be enhanced, and performance targets met. This was why a BIM focus was a starting point. Also, the case study building was kitted with state-of-the-art smart systems generating large amounts of energy performance data that provided valuable learning opportunities. The amount of information available at the stages of research commencement provided a rich source of data from which a BIM could be built, even when

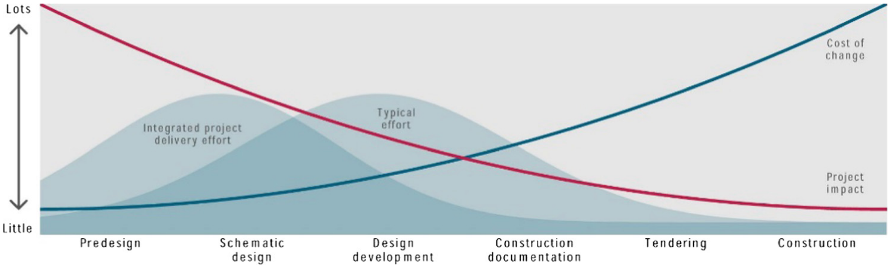


Fig. 4. MacLeamy curve, representing the effort required to change construction design per stage of design development [17]

the majority of this information was stored in formats not directly compatible with transfer to a BIM environment without significant effort. The impact this has on building design development was identified by MacLeamy [17], with Fig. 4 demonstrating how BIM is changing the information creation process.

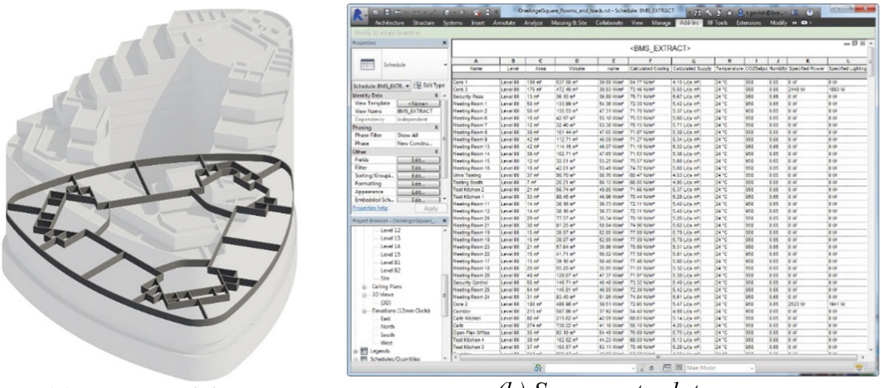
During the development of a BIM in which to store building energy performance information, methods of creating and transferring information within and outside the Revit platform were attempted. Lee et al. [18], Kim et al. [19], and Yalcinkaya and Singh [20] all identified the potential for error inclusion in this process, with many of these errors encountered, including:

- Interpretation of space bounding elements was variable across modelling tools. EPM tools require simplified geometry, but methods of simplifying this geometry are not yet available to the accuracy required for appropriate simulation use;
- Surface orientation necessary for EPM not inferred correctly;
- Intersections and slivers created as a result of poor geometry interpretation; and
- Data attribution lost between BIM authoring tools (geometric spaces ceased to be associated with their space meta-data).

In addition to user error, the remaining faults in data transfer came from incompatibility of data handling across modelling platforms, where interpretation and storage of information may differ significantly between tools, resulting in incompatible representation of the others original information [3, 21]. It was therefore determined that the BIM environment should be simplified similar to the building energy performance model, to provide a platform in which data could be attributed, without excess complexity of modelling preventing any modifications later made to support the research. Spaces monitored by the on-site BMS were modelled (comprising all inhabited areas within the building and plant and service spaces without regular occupancy), with their design performance attributes provided by the energy performance model to provide a performance describing BIM.

5 Output

The simplified BIM which was used as the basis for data attribution was manually rebuilt using the process described in Sect. 4.4 to avoid errors in data translation, and contained basic geometry describing the building's form, with space meta-data describing purpose and predicted performance (from IES-VE) for comparison with operational data when available (Fig. 5). The impact of simplification of the building modelled against the as-built building had little effect on the overall outcome of this research (Sect. 6), given the purpose of that simplification and granularity of data collected in the as-built building. However, the need for this simplification was indicative of the extent of computing power, memory and data storage ontologies required to support the processes demonstrated in Sect. 4.3, necessitating more structured and efficient means of handling descriptive building performance information.



(a) Space model

(b) Space meta-data

Fig. 5. Simplified partial BIM containing only spaces and their characteristic performances

5.1 BIM and Performance Meta-data Extraction

Data extraction from a BIM environment often requires access to the model via the same platform in which that model was created [22]. Codinhoto et al. [23] identified that access to data stored within BIM environments is a factor in reducing its adoption by Facilities Management (FM). Accessibility has increased through development of tools interoperating between BIM authoring platforms [24]; however, a gap between the data generated during design and use remains that could be overcome using basic data management techniques. Because of the proliferation of open exchange formats such as Industry Foundation Classes (IFC), information can be accessed via less costly viewing tools and open source alternatives.

A Dynamo [25] script extracting space geometry information in conjunction with related spatial performance meta-data of the case-study building was created [26]. This information was extracted from the Revit BIM environment into a JavaScript Object

Notation (JSON) lightweight data-interchange format (an object-mappable file which could be queried and accessed without need for proprietary software) capable of interpretation via the development language used. Utilising a non-standard format for extracting and processing data from the BIM environment distinguished the non-platform specific barriers to wider implementation of BIM from its authoring software. Dynamo was also used to attribute predicted performance data to the design model as meta-data describing spatial and system performance (Fig. 6). The more widely used IFC format was also considered as an appropriate carrier for this information but given the limitations in extract from Revit into this format and potential loss of data [27], the alternative was created to avoid these errors and specify exact data to be included in output of a lightweight and platform agnostic format.

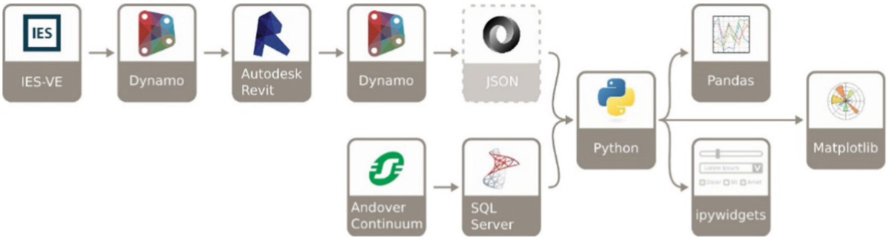


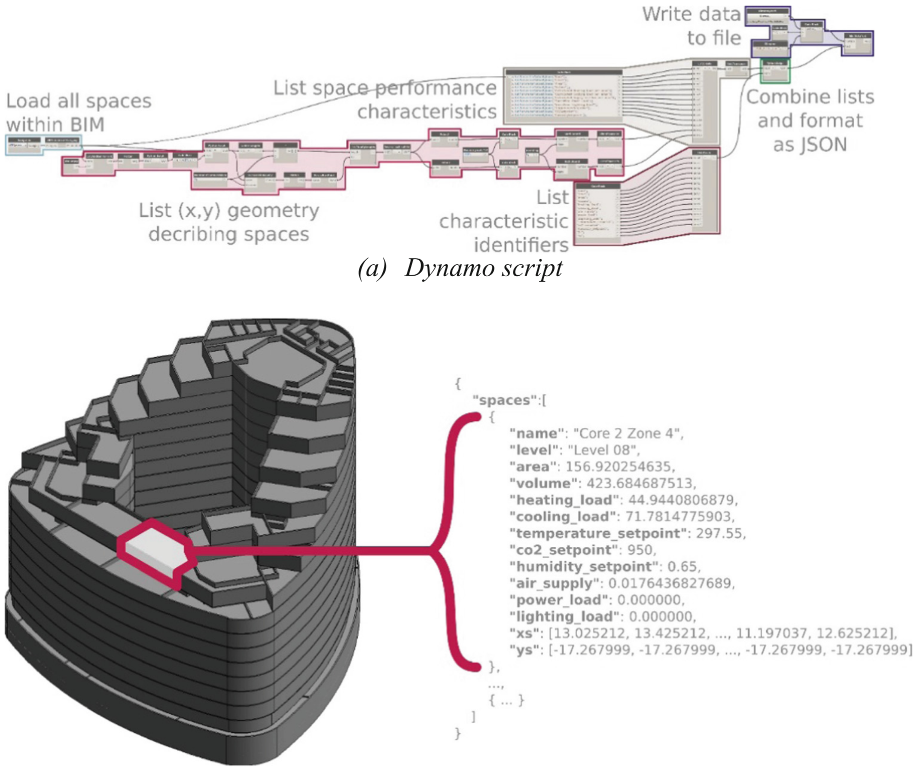
Fig. 6. Data transfer, extraction and visualisation process

Use of these tools suggest a potential change in the role of the engineer in this process, as applied programming requires knowledge of the purpose of that programming, and where creation of scripts automating engineering processes must account for the needs of the engineer and the task. Khaja et al. [24] and Fan et al. [28] suggest that the skills necessary for these processes are becoming more common yet lag behind the pace of development of tools in this area. The potential for handling of large amounts of information in this way also suggests the platforms commonly used to design, organise and access engineering related data may not be suitable [28, 29], and demonstrate a need for new tools to assist in this new data paradigm. The Dynamo script used to extract space performance meta-data from the simplified BIM shown in Fig. 7, outputting a JSON file interpretable outside any proprietary BIM authoring environment.

At this stage of research, the method used to interact with this data was undefined; therefore accessibility of data was essential to support later development of BIM and building performance linking tools (further justification for this choice and description of the data extraction and interaction process given in Sect. 6).

5.2 Summary and Findings

Collation of this data into a set of parametrically rich models created an environment from which performance data could be extracted, linked and utilised for later investigation of BIM as a performance management tool. Several findings were made



(a) Dynamo script

```
{
  "spaces":[
    {
      "name": "Core 2 Zone 4",
      "level": "Level 08",
      "area": 156.920254635,
      "volume": 423.684687513,
      "heating_load": 44.9440806879,
      "cooling_load": 71.7814775903,
      "temperature_setpoint": 297.55,
      "co2_setpoint": 950,
      "humidity_setpoint": 0.65,
      "air_supply": 0.0176436827689,
      "power_load": 0.000000,
      "lighting_load": 0.000000,
      "xs": [13.025212, 13.425212, ..., 11.197037, 12.625212],
      "ys": [-17.267999, -17.267999, ..., -17.267999, -17.267999]
    },
    ...,
    { ... }
  ]
}
```

(b) Revit data extract

Fig. 7. Dynamo script with psuedocode annotation describing BIM data extraction process

throughout the collection and utilisation of data for creation of representative models, the conclusions of which include:

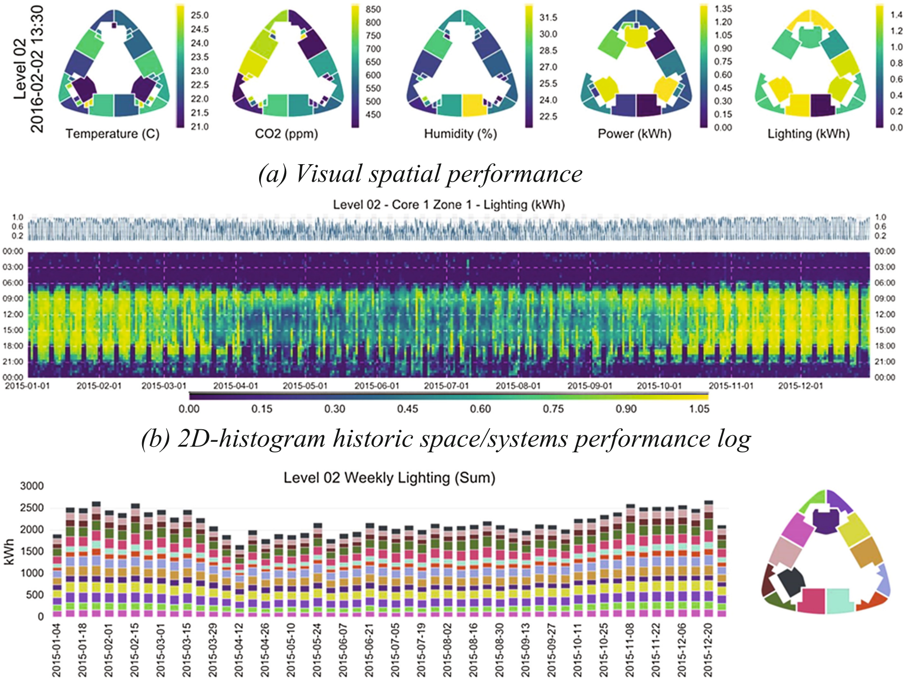
- Upon creation of handover documentation (in O&M manuals and related drawings, guides and specifications), information describing the building is already out of date. Changes made during commissioning of the building may not be reflected in documentation, nor changes made during occupation where occupant behaviour and use of space can vary significantly from design specifications [30, 31]. The result of these changes can be incorrect operation of conditioning systems to conditions no longer required, excess energy consumption through inefficient operation of plant equipment and discrepancies between the building its operational documents resulting in slower fault finding and fixing by FM;
- Requirements for simulation models do not translate well from BIM. These include the quality of space bounding [32], interpretation of meta-data between simulation tools [14] and level of detail suitable for inclusion in each environment [21];
- A large amount of information is being duplicated and superseded using traditional documentation methods (Sect. 4.1). Until revision control as part of BIM implementation can be implemented as a standard working process, the inclusion of

superseded or incorrect documents in ongoing design development is likely to continue, furthered by the utilisation of multiple design development platforms outside federated and integrated modelling environments [33]; and

- The data extracted from the embedded smart energy systems was largely numeric in nature, representing different metrics (e.g. lighting in kWh, CO_2 in ppm, and temperature in °C). It is a challenge to interpret meaning from this amount of information or identify trends 'hidden' within this voluminous dataset without predefined methods for processing or summary, if sense is to be made from it and the learning extracted.

From research outset, it was evident that the volume of information being generated through build monitoring, presented 'sense-making' challenges using conventional graphical and tabular means. Thus, it was thought appropriate to visualise the data to cut out the 'noise' and observe the data trends visually. These visualisation techniques mapped the performance and helped with understanding the performance 'behaviour pattern'. If conforming to the norm, the values would typically fall within a normal expected performance boundary. Outliers would be indicators of a system malfunction, its overuse and underuse due to unusual usage of building space/s (e.g. large number of visitors, doors windows being left open). This 'visual performance map', presents opportunities to develop intervening strategies that prompt 'remedial' actions so that the expected boundary conditions are maintained, lessons captured and performance optimised.

This approach to visualise large volumes of data describing building performance is novel and presents 'learning' opportunities for the future. With time as a 'critical mass' of data about buildings and their performance behaviours is accumulated, so will the opportunity to maintain 'healthy' buildings grow. These are the next big steps for industry to avail of and learn from. Examples of some of the visual performance trends are included in Fig. 8. showing where metering was incorrectly installed, thus not reporting values, where further calibration of sensors may be required and allowing the viewer to immediately recognise performance trends and outliers. For a building with 1000 sensors, each monitoring at 1-s intervals and storing values in an SQL database in float format (at 8-bytes each), this would result in 691.2 megabytes being generated each day. Because of the magnitude of the data, the first and foremost challenge was to represent the data in a simple format so that 'trends' could be observed, then classified and learning opportunities exploited. For example, Fig. 8b shows annual lighting load for a single zone in the case-study building, clearly showing where it is automatically dimmed during the summer months where more natural light is available, and where occupant sensing is present. Interesting patterns visible include the operation of lights overnight corresponding to security walk rounds, changes between days of week for operational periods and indication that on some days the daylight dimming is not working (later discovered to be due to occupants leaving blinds down preventing the light sensors from reducing the artificial lighting levels).



(a) Visual spatial performance

(b) 2D-histogram historic space/systems performance log

(c) Summary performance record

Fig. 8. Data visualisation from a BIM and BMS records using Python Pandas and Matplotlib

6 Method for the Management of Building Performance Data Using Design Data

A method to link predicted performance data with and monitored data captured during the occupation of the case-study building was developed by applying the findings determined through the previous tasks. The aim being to develop a prototype for the interconnection of these two distinct, yet related sources of performance information using existing technologies. A triangulation approach using a version of throwaway prototyping was applied to this task. This is the process of rapidly developing elements for incremental inclusion into a finished system. As no commercially applicable system was developed here (instead favouring the rapid development of a prototype), the composite elements created were used to evaluate the potential for, and challenges in implementing a system using BIM for performance analysis and management, demonstrating integration of the systems necessary for BIM performance management [34].

6.1 Extraction from the BIM Environment

The BIM environment to which data would be related was created in Sect. 5, where simplification of the as-designed BIM environment was used to generate a basic representation of the building as the BMS understands it. Extraction of static design information held in that environment into a lightweight, platform independent attribute-variable format (JSON) provided a means of accessing such information without the need for proprietary software the building users may not have.

JSON was chosen due to the researcher's familiarity with the format, its human interpretable structure and extensive support for parsing by multiple programming languages. Alternative formats are available, including IFC; however, in earlier investigations it was found that the existing data attribution capabilities of IFC for extensible meta-data attribute storage was limited and could potentially result in inaccessible or poorly structured data within the building model [3]. Storage in a related Binary JavaScript Object Notation (BSON) format was considered, utilising a MongoDB database [35] method of data storage; however, given the requirements for speed and portability in developing the throwaway prototype HDF5 was chosen as a storage format for the monitored performance data instead. This method of structuring large datasets in hierarchical data tables indexed using timestamps provided and highly responsive method of accessing and processing descriptive time-series performance data.

6.2 Making BIM Data Accessible

Data provided upon building handover is usually held in conventional formats such as spreadsheets, documents and drawings. This secondary data, while useful for quick interpretation and extraction of meaning, does not easily support further processing due to the limits imposed upon it by the processing already undertaken. Pollock [36] suggests the deficits to portrayal of information in this way include restrictions on access, reliance on interface-centric rather than data-centric views of information and undue effort placed on formatting of the usable data both by designers and processors, potentially limiting actions ensuring the accuracy and availability of all supporting information.

Utilising design-stage building energy performance data is contingent on its availability, accuracy and usability in a form manageable by the applied tools and methods. A key factor allowing attribution of building performance information to spaces and systems are comparable objects to which that data can be linked. Attribution of data to an object representing one of these elements must utilise an identifier distinguishing that element from others, relatable between models and datasets. This was achieved here by using common space and system names between the BIM and BMS datasets, but could be replicated with adherence to naming conventions and creation of dictionaries relating disparate yet related datasets where commonality is unavailable. Script 1 shows the JSON format used as the carrier for design-based BIM data, for connection with times-series performance information from the BMS. The file this represents was created using the Dynamo script shown in Fig. 7, containing basic information, constituting an as-designed description of the Revit models spatial composition and performance characteristics.

```
{
    "spaces":[
        {
            "name": "Core 2 Zone 4",
             "level": "Level 08",
            "area": 156.920254635,
            "volume": 423.684687513,
            "heating_load": 44.9440806879,
            "cooling_load": 71.7814775903,
            "temperature_setpoint": 297.55,
            "co2_setpoint": 950,
            "humidity_setpoint": 0.65,
            "air_supply": 0.0176436827689,
            "power_load": 0.000000,
            "lighting_load": 0.000000,
            "xs": [13.025212, 13.425212, ..., 11.197037, 12.625212],
            "ys": [-17.267999, -17.267999, ..., -17.267999, -17.267999]
        },
    ]
}
```

Script 1: Example JSON format space object characteristic extracted from Autodesk Revit using Dynamo

The processing required to create the datasets supporting the link between data in a BIM environment and in performance design and monitoring systems requires skills in areas which designers and operators may not possess. Automation of these processes would be necessary for implementation in a wider range of projects, for which standardisation of procedures and design documentation would be required. Existing standards detailing the naming, storage and data handling methods in and around BIM environments (such as the Data Dictionary provided by BuildingSMART [37] and BSI [38] on which it is based), would provide a good starting point from which automation could be developed.

6.3 Data Relation

Kohlhase [39], Thorne and Ball [40], Chen and Chan [41], and Hendry and Green [42] identify the limitations of data portrayal as it implemented in a BMS currently, with visualisation of the data being collected an integral part to the tools developed here. Those limitations include speed of access, interpretation and action through ineffective information structuring, relying on user familiarity with the document rather than self-documented logical data structuring such as that shown in Script 1. Following sourcing, extraction and processing of data, a means of accessing both the BIM (as a JSON file) and BMS (as a HDF5 file) was developed. The need for efficient handling of time-series performance data collected by the BMS and sensor network throughout the case-study building was essential, given the intractability of monitored data and requirement for ease of interpretation by building operators in identifying performance trends and opportunities for improvement.

The existing means of querying data from the BMS was inefficient due to the lack of indexing applied to collected data in the SQL environment [26], and would be an inhibiting factor in the portrayal of performance data linked to the BIM in other buildings. This was implemented in the case-study building without accessibility to information by FM without supervision by the providers of the BMS software, significantly increasing the time taken to identify performance deficiencies and trends. The solution developed was based upon static representation of the as-designed building and its historical performance up to the point at which extraction of such data is made from the BMS; however, there is potential for a link between a live representation of the as-managed building as both a descriptive model and monitored performance, given efficient access to this data and the continuous update of a representative model.

The tools used in the development of a method for linking BIM and performance monitoring are indicated in Table 2. Sources of information for this process are typical of commonly used industry standard software, supplemented by the programming language used in the case-study (Python) and supporting packages included as the means through which data interoperability and interpretation was achieved between the two environments.

Table 2. Software used during development of the BIM-linked performance monitoring method

Software	Function
IES-VE [15]	Modelling and simulation of building performance
Autodesk Revit [43]	Modelling and attribution of descriptive performance meta-data to objects and spaces in a BIM environment
Autodesk Dynamo [25]	Extraction of geometry and meta-data from Autodesk Revit into a lightweight data-interchange format (JSON)
Andover Continuum Cyberstation	Front-end interface to BMS
SQL Server 2008	Back-end BMS storage of historic performance data
Python	
Pandas [44]	Extraction of data from SQL Server, cleaning of extracted data and code to interlink JSON file with queryable HDF5 performance data store
Matplotlib [45] Ipywidgets [46]	Visualisation of performance data and user interaction elements

Data Relation Process

Figure 9 illustrates the process followed in gathering and linking the data contained within the distinct datasets, associating data from the BMS to objects within the BIM without specification of distinct software. These actions represent high-level processes by which the data is generated, collected and utilised from the prediction of building energy performance to its storage in a BIM environment, and connection to monitored performance from a BMS. A prescriptive methodology is unsuitable for the wider

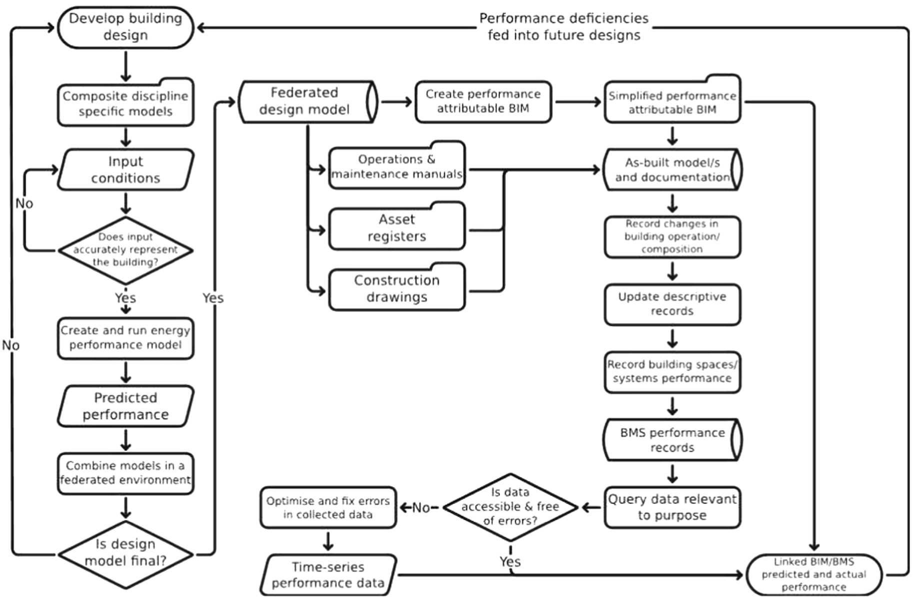

Develop building design
Performance deficiencies fed into future designs
Composite discipline specific models
Input conditions
No
Does input accurately represent the building?
Yes
Create and run energy performance model
Predicted performance
Combine models in a federated environment
Is design model final?
No
Yes
Federated design model
Create performance attributable BIM
Simplified performance attributable BIM
Operations & maintenance manuals
Asset registers
Construction drawings
As-built model/s and documentation
Record changes in building operation/ composition
Update descriptive records
Record building spaces/ systems performance
BMS performance records
Query data relevant to purpose
Is data accessible & free of errors?
No
Optimise and fix errors in collected data
Time-series performance data
Yes
Linked BIM/BMS predicted and actual performance

Fig. 9. BIM/performance data information flow and linking process

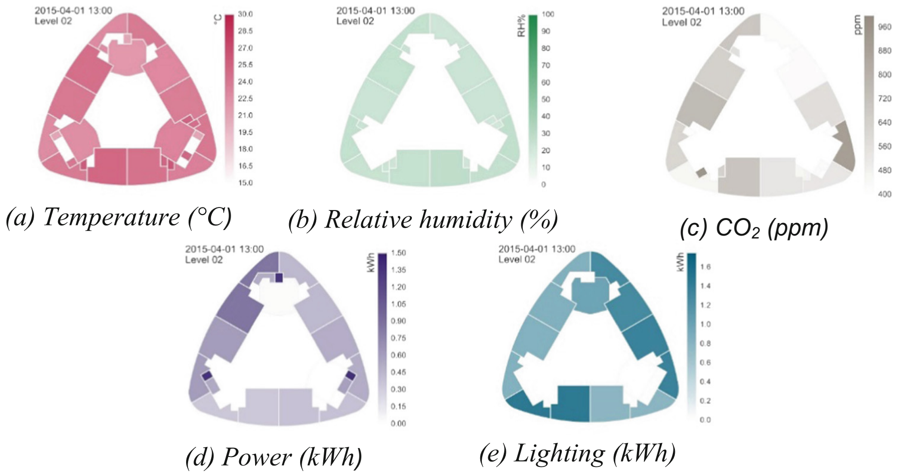
industry given the non-homogeneity of design and operation methods, tools and processes, and the need for implementation considering the needs of each individual building project [26].

Data Portrayal

The purpose of linking design and operation data has been to provide a method of performance interpretation for those responsible for occupying, operating and managing the performance of that building. The following tools supported by the BIM/BMS link developed here are described, indicating the capabilities of such a system and its potential for BIM supporting performance management through basic interpretation and connection of data using an efficient, open and accessible method.

Space Attributes

The monitored spatial performance descriptors of CO_2 levels, temperature, humidity, power and lighting energy consumption are attributed to the geometry extracted from the accessible JSON format and interpreted via Python. Quick visualisation of spatial performance in a floor-plan enables the operator to identify areas of performance deficiency to focus efforts on remediation and optimisation. A snapshot in time for Level 02 is shown in Fig. 10 showing spaces and their individual monitored variables. Several spaces lighting and small power monitoring are not available, indicating potential errors in the sensors or BMS monitoring these.



(a) Temperature (°C) *(b) Relative humidity (%)* (c) CO_2 (ppm)

(d) Power (kWh) *(e) Lighting (kWh)*

Fig. 10. Space attributes showing ‘snapshot’ performance characteristics

2D-Histogram of Historical Performance

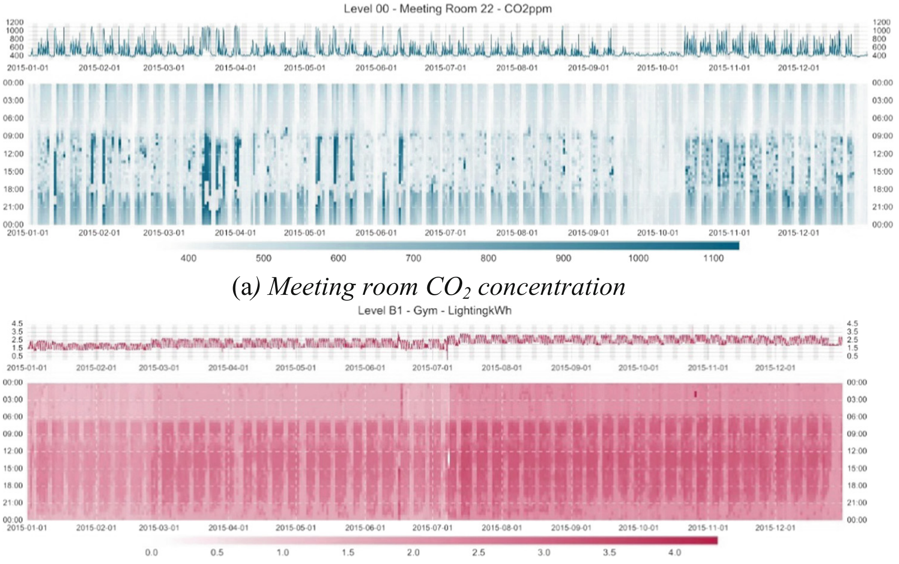
Historic portrayal of performance in a 2D-histogram format has been demonstrated by Yarbrough et al. [47] and Meyers et al. [48] as providing a suitable means of efficiently displaying large amounts of time-series data. Application to the data collected show some significant trends and opportunities for improvement in the management of the case-study building. Spatial performance characteristics shown in Fig. 11 show how occupant behaviour can be inferred from monitored performance, where monitoring is implemented correctly. Periods where the meeting room described in Fig. 11a is occupied can be clearly seen as increases in local CO2 levels, with the space identified as. unoccupied for 68% of the time during occupied hours[1]. The rate of air change can also be compared against external CO2 levels; as the building is vacated at the end of the day and ventilation systems turn off, the amount of ambient CO2 in the air spikes around 20:00 and returns to external ambient conditions.

A trend towards less efficient performance can be identified in Fig. 11b, with a 23% increase in energy used for lighting between the first and second halves of the year following the change in operational hours from 06:30-23:30, to 24-h use. Lights should turn off automatically during unoccupied hours which are not happening as indicated by the 2.2 kWh base load overnight following the change (a 49% increase in unoccupied lighting loads).

Performance Summary

Summarising the energy consumed by distinct spaces within the building is useful to the FM and estates management team to understand where energy has been used, and how each metered space compares to identify opportunities for improvement. Following data cleansing and storage in an HDF5 file, the process used to query data and

[1] During 2015 and between 08:00 and 18:30, 2343 out of a possible 3443 h showed CO_2 levels within 10% of the external ambient CO_2 level.



(a*) Meeting room CO_2 concentration*

(b) Gym lighting power consumption

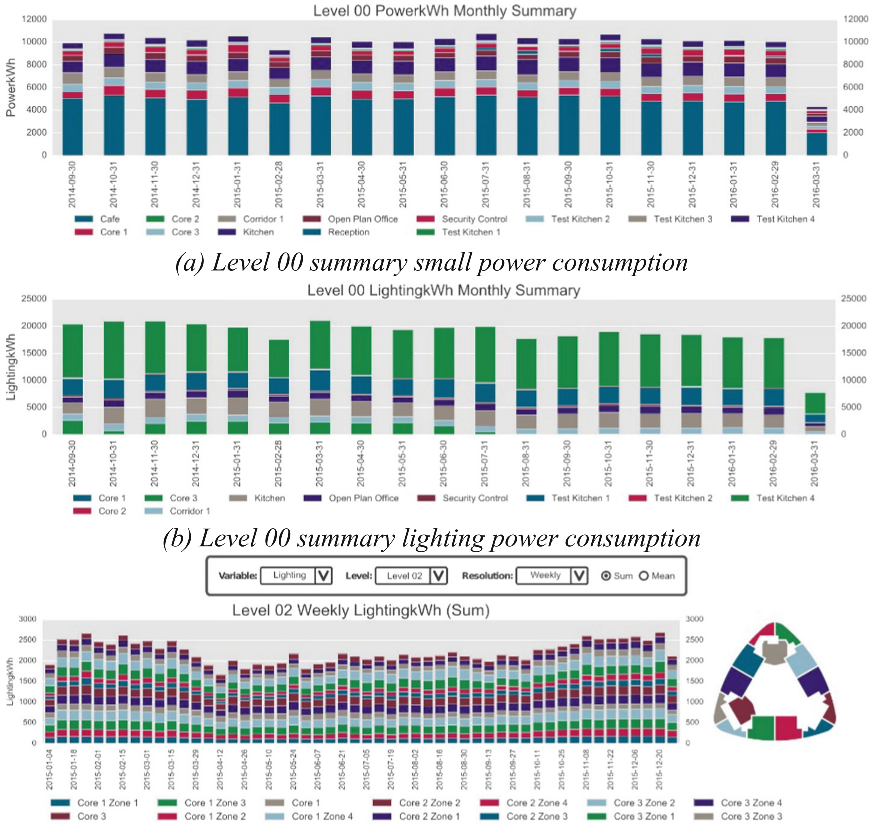
Fig. 11. Time-series plot and heat-map spatial performance visualisation

create Fig. 12 took seconds rather than the hours required for extraction from the un-optimised BMS SQL database, demonstrating the room for improvement in this process. Using Pandas [44], analysis resolution can be easily adjusted to show more granular detail, showing the effects of holidays and the daylight dimming in-place across the floor analysed (Fig. 12c), with user interaction modifying summary parameters to explore all aspects of the building's spatial performance.

The trends expected from such data, in-line with the patterns of use and response to external and internal climate factors are classified as multiplicative, resulting from those factors generated from differential responses to white noise inputs. As such, analysis of this data requires a combination of approaches to account for the variability between predictable (time-of-day, day-of-week, season-of-year) and unpredictable (occupant behaviour, system operation issues and unexpected) influencers. Therefore, the data obtained corresponds to periodic and sinusoidal variations, oscillating according to diurnal, weekly and seasonal differences [49].

Aggregation for Diurnal Trend Analysis

As the amount of data made available to FM increases, the opportunities for trend analysis of operational profiles increase correspondingly, with access to many data points from which to draw aggregated profiles of operation. Figure 13 demonstrates this, showing how water use by the whole building varies per season and weekday, and signifying the average setback consumption outside occupied hours. While not strictly BIM application to performance management, the processes followed to enable access to information efficiently to support BIM integration, forces the monitored data to be efficiently structured, enabling analysis extemporaneous to conventional summation and averaging [2].



(a) Level 00 summary small power consumption

(b) Level 00 summary lighting power consumption

Variable: Lighting
Level: Level 02
Resolution: Weekly
● Sum ○ Mean

(c) User interactive performance summary

Fig. 12. Case-study building lighting and small power summaries

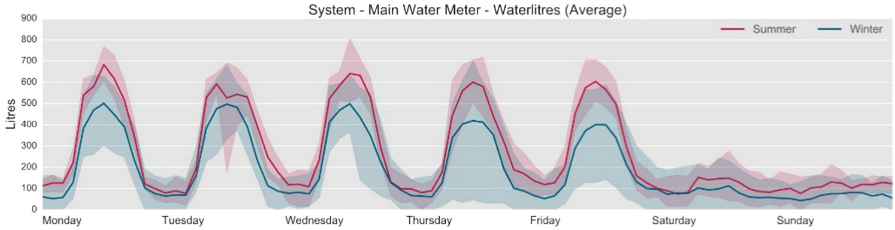


Fig. 13. Aggregated mean diurnal profiles based on day of week and time of year

Predicted Performance Disparity Indication

The primary means of distinguishing performance disparity between the predicted and monitored building using the BIM and BMS data sources is achieved via the creation of a ‘performance dashboard’ using Python. This uses the set-points defined within the JSON BIM representation, in conjunction with the data collected via the BMS to indicate levels of performance of the operational building compared with these. Figure 14 shows a snapshot of this dashboard. Many meters are non-reporting (due to commissioning issues), indicating significant room for improvements in installation and commissioning of the sensors network and metering system. Spaces at above the specified maximal operating conditions are indicated for attention of the building operator.

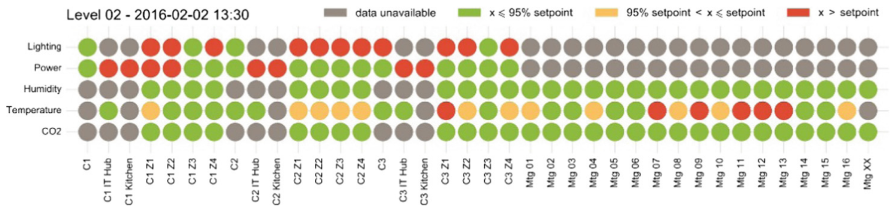


Fig. 14. Snapshot of the dashboard with interactive settings to override and adjust sensitivity settings for the indicators

7 Validity and Application

The case-study approach and wide variability across the AEC industry means generalisation of research findings is difficult. The findings presented here are based on research methods developed with consideration of their applicability to the processes being examined, with those methods employed to generate widely applicable findings. Reliability of the data used is dependent on the systems in place recording that data describing the building being monitored; however, the processes utilising that information presented here may be applied to other non-domestic buildings and are not specific to the case-study. The following issues were noted with the data collected here:

- Data collected from the case-study building’s BMS was processed to remove errors, potentially reducing its accuracy;
- Spatial performance attributes may not be attributed to the correct spaces in the tool demonstrating a BIM and performance data link due to the BMSs lack of structure at the point of data extraction; however, this does not impact findings; and
- The changeable design environment in which BIM is applied means replication of the processes detailed here on other projects may be difficult. However, care was taken to avoid specification of methodologies relevant only to the case-study used, and conclusions made relevant to the wider construction industry.

8 Discussion and Conclusions

The move towards the use of systems to record data about the performance of building presents an opportunity to examine patterns of behaviour, which shed light on the performance peaks and troughs. Answers often lie within these peaks and troughs. Each system is designed for an expected 'behaviour pattern' within a 'threshold' boundary. This is the expected performance the system is designed for. Any deviations in the performance are indicators of a system malfunction, its overuse or underuse, and identification of these is essential for continuous optimisation of building performance and operation.

Performance when bounded within the threshold limits would be considered to be 'normal' or 'expected'. Thus, answers fall within four categories of performance behaviours. These include system malfunction, system overuse, system underuse, and normal performance. Simple classification of these enables grouping of identified trends; however, automated identification would require more in-depth learning using larger datasets, machine-learning techniques and expertise of building operators to categorise patterns across a hugely variable data landscape.

Data being collected over a period of time presents opportunities to observe system behaviour patterns and map those patterns for classification within learning clusters of 'expected' and 'unexpected' thresholds. Doing so would enable building owners to truly understand the building, so performance can be firstly quantified and then optimised. Thus, an unusual activity that is significantly beyond the expected 'norm' would present an opportunity to learn about the building so a healthy function can be determined and maintained.

Numeric data sets represented as spreadsheets present challenges and require application of complex algorithms to sift through data and subsequently sense make. Generation of large datasets across multiple buildings through extensive monitoring has created the potential for application of big-data analytics for holistic performance assessment and pattern recognition. In order to make the most of this plentiful source of data, its management during design and operation must be considered. This paper demonstrated how large volumes of performance data could be organised visually, so performance trends are first observed, queried and classified, and then lessons learnt are fed back for further performance enhancement. This when combined with the development of a standard information structure and a measured building performance data ontology would negate the need to manually process information prior to analysis and reduce steps necessary to interpreting patterns and identifying trends. Application of techniques used in the IT sector, and adherence to common standards could make this possible; however, the skills necessary to implement this can be a barrier to its adoption. For this gap to be met, a new breed of 'building' researcher and/or consultant with a complex skillset is needed. They who would be adept at not only understanding (and subsequently interpreting) the design considerations and performance requirements of clients but would also be adept at programming and navigating through multiple environments, such as those outlined in Fig. 6). This would ensure that the vast volume of data extracted from complex systems is firstly organised and classified in a meaningful way, relevant queries are determined, trends are observed, and visual representation techniques used to derive new meaning and determine new classifications.

References

1. Gerrish, T., Cook, M.J., Ruikar, K.: BIM for the management of building services information during building design and use. Sci. Technol. Built Environ. **22**(3), 249–251 (2016). https://doi.org/10.1080/23744731.2016.1156947
2. Gerrish, T., Ruikar, K., Cook, M.J., Johnson, M., Phillip, M.: Analysis of basic building performance data for identification of performance issues. Facilities **35**(13/14), 801–817 (2016). https://doi.org/10.1108/f-01-2016-0003
3. Gerrish, T., Ruikar, K., Cook, M.J., Johnson, M., Phillip, M.: Attributing in-use building performance data to an as-built building information model for life-cycle building performance management. In: Proceedings of CIB W78, 27–29 October 2015. CIB, Eindhoven, The Netherlands (2015)
4. Gerrish, T., Ruikar, K., Cook, M.J.: Cross discipline knowledge transfer for concurrent BIM adoption in an engineering organisation. In: Proceedings of the 2014 CIB W55/65/89/92/96/102/117 & TG72/81/83 International Conference on Construction in a Changing World, 4–7 May 2014. CIB, Kandalama, Sri Lanka (2014)
5. RIBA: RIBA Plan of Work 2013 Overview. Report. RIBA, London, UK (2013)
6. Bowen, G.A.: Document analysis as a qualitative research method. Qual. Res. J. **9**(2), 27–40 (2009). https://doi.org/10.3316/QRJ0902027
7. Shermer, M.: How the survivor bias distorts reality (2014). http://www.scientificamerican.com/article/how-the-survivor-bias-distorts-reality/. Accessed 15 June 2016
8. Thabet, W., Lucas, J., Johnston, S.: A case study for improving BIM-FM handover for a large educational institution. In: Proceedings of the Construction Research Congress (CRC), 26–28 September 2016, pp. 2177–2186. American Society of Civil Engineers, San Juan, Puerto Rico (2016). https://doi.org/10.1061/9780784479827.217
9. Lucas, J., Bulbul, T.: An object-oriented model to support healthcare facility information management. Autom. Constr. **31**, 281–291 (2013). https://doi.org/10.1016/j.autcon.2012.12.0142
10. Hjelt, M., Björk, B.-C.: Experiences of EDM usage in construction projects. J. Inf. Technol. Constr. **11**, 113–125 (2006)
11. European Construction Research Network. E-CORE Strategy for Construction RTD (2005). http://www.e-core.org/strategy. Accessed 15 June 2016
12. Gallaher, M.P., O'Connor, A.C., Dettbarn, J.L., Gilday, L.T.: Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry. NIST GCR 04-867. National Institute of Standards and Technology, Gaithersburg, MD (2004)
13. Coakley, D., Raftery, P., Keane, M.: A review of methods to match building energy simulation models to measured data. Renew. Sustain. Energy Rev. **37**, 123–141 (2014). https://doi.org/10.1016/j.rser.2014.05.007
14. Bazjanac, V.: IFC BIM-based methodology for semi-automated building energy performance simulation. In: Proceedings of CIB W78, 15–17 July 2008, Santiago, Chile (2008)
15. Integrated Environmental Solutions. Virtual Environment 2015. Version 2015 (2016). https://www.iesve.com/software/ve-for-engineers. Accessed 1 Jan 2016
16. U.S. Department of Energy. International Performance Measurement & Verification Protocol. Concepts and Options for Determining Energy and Water Savings. U.S. Department of Energy (2002)

17. MacLeamy, P.: The Future of the Building Industry: The Effort Curve. HOK (2010). https://youtu.be/9bUlBYc_Gl4. Accessed 14 June 2016
18. Lee, S.-K., Kim, K.-R., Yu, J.-H.: BIM and ontology-based approach for building cost estimation. Autom. Constr. **41**, 96–105 (2014). https://doi.org/10.1016/j.autcon.2013.10.020
19. Kim, H., Anderson, K., Lee, S.-H., Hildreth, J.: Generating construction schedules through automatic data extraction using open BIM (building information modeling) technology. Autom. Constr. **35**, 285–295 (2013). https://doi.org/10.1016/j.autcon.2013.05.020
20. Yalcinkaya, M., Singh, V.: Patterns and trends in building information modeling (BIM) research: a latent semantic analysis. Autom. Constr. **59**, 68–80 (2015). https://doi.org/10.1016/j.autcon.2015.07.012
21. Gerrish, T., Ruikar, K., Cook, M.J., Johnson, M., Phillip, M.: Using BIM capabilities to improve existing building energy modelling practices. Eng. Constr. Archit. Manag. **24**(2) (2016). https://doi.org/10.1108/ecam-11-2015-0181
22. Aranda-Mena, G., Wakefield, R.: Interoperability of building information – myth or reality? In: Martinez, M., Scherer, R. (eds.) eWork and eBusiness in Architecture, Engineering and Construction. Proceedings of the European Conference on Product and Process Modeling (ECPPM), 13–15 September 2006, Valencia, Spain, pp. 127–133. Taylor & Francis Group (2006)
23. Codinhoto, R., Kiviniemi, A., Kemmer, S., Essiet, U.M., Donato, V., Tonso, L.G.: BIM-FM implementation: an exploratory investigation. Int. J. 3-D Inf. Model. **2**(15), 1–15 (2013). https://doi.org/10.4018/ij3dim.2013040101
24. Khaja, M., Seo, D.J., McArthur, J.J.: Optimizing BIM metadata manipulation using parametric tools. Procedia Eng. **145**, 259–266 (2016). https://doi.org/10.1016/j.proeng.2016.04.072
25. Autodesk: Dynamo. Version 0.9.0 (2015). http://dynamobim.org. Accessed 3 Mar 2016
26. Gerrish, T., Ruikar, K., Cook, M.J., Johnson, M., Phillip, M., Lowry, C.: BIM application to building performance visualisation and management: challenges and potential. Energy Build. **144** (2017). https://doi.org/10.1016/j.enbuild.2017.03.032
27. Solihin, W., Eastman, C., Lee, Y.-C.: Toward robust and quantifiable automated IFC quality validation. Adv. Eng. Inform. **29**(3), 739–756 (2015). https://doi.org/10.1016/j.aei.2015.07.006
28. Fan, C., Xiao, F., Madsen, H., Wang, D.: Temporal knowledge discovery in big BAS data for building energy management. Energy Build. **109**, 75–89 (2015). https://doi.org/10.1016/j.enbuild.2015.09.060
29. Rathore, M.M., Ahmad, A., Paul, A., Rho, S.: Urban planning and building smart cities based on the Internet of Things using big data analytics. Comput. Netw. **101**, 63–80 (2016). https://doi.org/10.1016/j.comnet.2015.12.023
30. Wolfe, A.K., Malone, E.L., Heerwagen, J., Dion, J.: Behavioral Change and Building Performance: Strategies for Significant, Persistent, and Measurable Institutional Change. Report. Pacific Northwest National Laboratory (2014)
31. Clevenger, C.M., Haymaker, J.: The impact of the building occupant on energy modeling simulations. In: Proceedings of the International Conference on Computing in Civil and Building Engineering (ICCCBE), 14–16 June 2006, Montréal, Canada, pp. 1–10 (2006)
32. Bazjanac, V.: Space boundary requirements for modeling of building geometry for energy and other performance simulation. In: Proceedings of CIB W78, 16–18 November 2010, Cairo, Egypt (2010)
33. Dubler, C.R., Messner, J.I., Anumba, C.J.: Using lean theory to identify waste associated with information exchanges on a building project. In: Proceedings of the Construction Research Congress (CRC), 8–10 May 2010, pp. 708–716. American Society of Civil Engineers, Alberta, Canada (2010). https://doi.org/10.1061/41109(373)71

34. Korpela, J., Miettinen, R., Salmikivi, T., Ihalainen, J.: The challenges and potentials of utilizing building information modelling in facility management: the case of the center for properties and facilities of the University of Helsinki. Constr. Manag. Econ. **33**(1), 3–17 (2015). https://doi.org/10.1080/01446193.2015.1016540
35. MongoDB, Inc.: MongoDB 3.2. MongoDB, Inc. (2016). https://docs.mongodb.com/manual. Accessed 26 Aug 2016
36. Pollock, R.: Give Us the Data Raw, and Give it to Us Now (2007). http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now. Accessed 3 July 2016
37. BuildingSMART: BuildingSMART Data Dictionary (2016). http://bsdd.buildingsmart.org. Accessed 3 July 2016
38. BSI: ISO 12006-3:2007. Building construction – Organization of information about construction works – Part 3: Framework for object-oriented information. British Standards Institution (2007)
39. Kohlhase, A.: Human-spreadsheet interaction. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part IV. LNCS, vol. 8120, pp. 571–578. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40498-6_47
40. Thorne, S., Ball, D.: Exploring human factors in spreadsheet development. In: Proceedings of the European Spreadsheet Risks Interest Group, 28 July 2005. EuSpRIG, Greenwich, UK (2005)
41. Chen, Y., Chan, H.C.: Visual checking of spreadsheets. In: Proceedings of the European Spreadsheet Risks Interest Group. EuSpRIG, Greenwich, London, July 2000
42. Hendry, D.G., Green, T.R.G.: Creating, comprehending and explaining spreadsheets: a cognitive interpretation of what discretionary users think of the spreadsheet model. Int. J. Hum Comput Stud. **40**(6), 1033–1065 (1994). https://doi.org/10.1006/ijhc.1994.1047
43. Autodesk: Revit. Version 2016 (2015). http://www.autodesk.co.uk/products/revit-family/overview. Accessed 3 Mar 2016
44. McKinney, W.: Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, June 28–July 3 2010, pp. 51–56 (2010). http://pandas.pydata.org. Accessed 1 Apr 2016
45. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
46. Pérez, F., Granger, B.E.: IPython: a system for interactive scientific computing. Comput. Sci. Eng. **9**(3), 21–29 (2007). https://doi.org/10.1109/MCSE.2007.53
47. Yarbrough, I., Sun, Q., Reeves, D.C., Hackman, K., Bennett, R., Henshel, D.S.: Visualizing building energy demand for building peak energy analysis. Energy Build. **91**, 10–15 (2015). https://doi.org/10.1016/j.enbuild.2014.11.052
48. Meyers, S., Mills, E., Chen, A., Demsetz, L.: Building data visualization for diagnostics. ASHRAE J. **38**(6), 63–72 (1996)
49. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications: With R Examples. Springer Texts in Statistics. Springer, New York (2006)
50. BuroHappold Engineering, Palin, T.: One Angel Square, Manchester (2016)

Author Index