

Computer Communications and Networks

Feng Chen

Rebeca I. García-Betances

Liming Chen

María Fernanda Cabrera-Umpiérrez

Chris Nugent *Editors*

# Smart Assisted Living

Toward An Open Smart-Home  
Infrastructure



Springer


# Computer Communications and Networks

## Series Editors

Jacek Rak, Department of Computer Communications, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland

A. J. Sammes, Cyber Security Centre, Faculty of Technology, De Montfort University, Leicester, UK

## Editorial Board

Burak Kantarci , School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada

Eiji Oki, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Adrian Popescu, Department of Computer Science and Engineering, Blekinge Institute of Technology, Karlskrona, Sweden

Gangxiang Shen, School of Electronic and Information Engineering, Soochow University, Suzhou, China

The **Computer Communications and Networks** series is a range of textbooks, monographs and handbooks. It sets out to provide students, researchers, and non-specialists alike with a sure grounding in current knowledge, together with comprehensible access to the latest developments in computer communications and networking.

Emphasis is placed on clear and explanatory styles that support a tutorial approach, so that even the most complex of topics is presented in a lucid and intelligible manner.

More information about this series at <http://www.springer.com/series/4198>

Feng Chen · Rebeca I. García-Betances ·  
Liming Chen · María Fernanda Cabrera-Umpiérrez ·  
Chris Nugent  
Editors

# Smart Assisted Living

Toward An Open Smart-Home Infrastructure

 Springer



*Editors*

Feng Chen  
School of Computer Science & Informatics  
De Montfort University  
Leicester, UK

Rebeca I. García-Betances  
Life Supporting Technologies  
Universidad Politecnica de Madrid  
Madrid, Spain

Liming Chen  
School of Computing  
Ulster University  
Belfast, UK

María Fernanda Cabrera-Umpiérrez  
Life Supporting Technologies  
Universidad Politecnica de Madrid  
Madrid, Spain

Chris Nugent  
School of Computing  
Ulster University  
Newtownabbey, UK

ISSN 1617-7975                      ISSN 2197-8433 (electronic)  
Computer Communications and Networks  
ISBN 978-3-030-25589-3              ISBN 978-3-030-25590-9 (eBook)  
<https://doi.org/10.1007/978-3-030-25590-9>

© Springer Nature Switzerland AG 2020

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, expressed 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.

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

# Preface

Assisted living in Smart Home (SH) can change the way of caring the older people and manage their conditions and maintain their well-being. This will support the ageing population to live longer independently and to enjoy the comfort and quality of life in their private environments. With the increasing ageing population and the growing demand on novel healthcare models, research on SH for independent living, self-management and well-being has intensified over the last decade coupled by the wide availability of affordable sensing and effective processing technologies. Yet it still remains a challenge to develop and deploy SH solutions that can handle everyday life situations and support a wide range of users and care applications.

The challenge to the rapid development and deployment of SH solutions essentially arises from the technology complexity and application diversity of the SH field. SH is a highly multidisciplinary research field involving a number of disciplinary areas and topics. To be successful any SH solution requires seamless technology integration and inputs from multiple subject areas. Researchers working in different technological disciplines usually have little understanding and appreciation of each other's research issues. There is little consideration of the "big picture", i.e. integration and interoperability. This leads to fragmented self-contained technologies, which are not suitable to serve as an integral part of a technology infrastructure to solve "bigger" complex problems. SH accommodates a wide range of applications, and each of them may require different sensors, data processing methods and intervention mechanisms. As technologies are developed in a specific application context, the resulting technology infrastructures are usually ad hoc, i.e. domain dependent, application specific, difficult to be applied to solve problems of a different application characteristic. This suggests that an alternative to a one-size-fits-all approach to develop SH technology infrastructure is needed in order to advance the state-of-the-art. SH technologies must be interoperable for seamless technology integration and rapid application development, and adaptable for easy deployment and management, achieved by thorough testing and validation in multiple application scenarios. In addition to multidisciplinary and application heterogeneity, SH is also a field involving multiple stakeholders, e.g. researchers, technology and solution developers, service providers, carers and end users.

Addressing the needs of a SH application solely from a single stakeholder's perspective is insufficient to deliver the right solution for the right users. Any best practices should be built upon effective communication and sharing of knowledge, and consensus of views and needs between stakeholders in the value chain.

As IoT industry advances, SH can leverage from cheap ubiquitous sensors, interconnected smart objects, packaged with robust context inference and interaction techniques so that SH technologies will be adaptive to fit versatile living environments, and interoperable for heterogeneous applications. In addition, the service-oriented cloud-based system architecture will support reconfiguration and modular design that is essential to empower care providers to customise their solutions. An easy-to-use open technology infrastructure which provides validated technology components and platforms built upon them is highly demanded. The technologies in the infrastructure should be modular and extensible and can be reused and automatically configured and integrated into a service infrastructure to facilitate wider adoption. By using this infrastructure, developers can rapidly develop functionality and applications, and care institutions and the elderly themselves could choose and configure solutions according to their needs.

This book is designed to make a critical contribution towards an open smart home (SH) technology infrastructure by interlinking disciplines from sensor technology and integration, context inferences, and interaction, to service infrastructures, and considering key principles of social impact, security and privacy. This book aims to be unique in its area because of the multidisciplinary integration scope that leads to the development of new effective, integrated, and interoperable SH solutions, taking into account multiple research areas. It offers killer applications for pervasive computing technologies, demonstrating and inspiring researchers in pervasive computing community how fundamental theories, models, algorithms can be exploited to solve real-world problems, thus impacting the society and economy. A total of 16 chapters are included in this book. These chapters are centred on different areas such as assisted living solutions, Smart Home (SH) user needs and system requirements, sensing and monitoring, activity recognition, context awareness, adaptive user interfaces, open SH infrastructures and toolsets. The main themes of the book are organised into four parts: (1) Sensing and Monitoring Technologies; (2) Activity Recognition and Behaviour Analysis; (3) User Needs and Personalisation, and (4) Open Smart Home and Service Infrastructures. The following briefly describes the chapters included in each part.

Part I addresses the issues of sensing and monitoring technologies. Chapter 1 reviews some state-of-the-art user context sensing techniques under smart home infrastructure. Chapter 2 describes a system that provides continuous localisation and behavioural analysis of a person's motion pattern over an indoor living space using multiple Kinect sensors. Chapter 3 proposes an unobtrusive sensing solution for monitoring post-stroke rehabilitation exercises within a home environment. Unobtrusive sensing solutions such as thermal, radar, optical and ultrasound are considered with practical examples. Chapter 4 explores how Google Glass can be used to annotate cystoscopy findings in a hands-free and reproducible manner by

surgeons during operations in the sterile environment inspired by the current practice of hand-drawn sketches.

Part II focuses on the context interference and behaviour analysis. Chapter 5 aims to accurately recognise different sports types in the Sports Video in the Wild (SVW) data-set employing transfer learning. The proposed system can be integrated with a smart home platform to identify sports activities of individuals and track their progress. Chapter 6 covers a study regarding the use of object detections as input for location and activity classification and analyses the influence of various detection parameters. Chapter 7 explores how the quality of data may affect the recognition performance. Outcomes are based on a comparison of activity recognition performance of six machine learning classifiers. Chapter 8 presents an automated screening approach to Prechtl's General Movement Assessment (GMA), based on body-worn accelerometers and a novel sensor data analysis method—Discriminative Pattern Discovery (DPD).

Part III concerns with personalisation and adaptive interaction. Chapter 9 presents experiences, best practices and lessons learned applying user-centred design methodology (UCD) in different European projects from several years of work conducted at LifeSTech group from UPM, in areas such as chronic diseases management, accessibility and cognitive rehabilitation. Chapter 10 introduces a theoretical framework on detecting user emotions during human–robot interaction and translating the detected user emotions into user mood estimates, enabling a service robot to adapt its assistive behaviour based on its user mood and emotions. Chapter 11 presents a semantic markup approach for the application of the Human Behaviour Monitoring and Support (HBMS) assistive system to achieve the required context awareness. The chapter shows how to semantically describe devices and web applications, and how personalised and adaptive HBMS user clients and the power of the context model of HBMS System can be used to bridge an existing activity recognition gap.

Part IV covers the framework and infrastructures of open smart home and service. Chapter 12 demonstrates a system that can turn a normal house to a smart house for daily activity monitoring with the use of ambient sensors. The multiresident activity recognition system is designed to support multiple occupants in a house with minimum impact on their living styles. Chapter 13 presents the living labs as the novel instruments for evaluating, assessing and validating innovative products, solutions or services in the particular domain of smart living environments. Chapter 14 proposes a cloud-based smart medical system by applying MapReduce distributed processing technology. A new distributed k-nearest neighbours (kNN) algorithm that combines the Voronoi inverted grid (VIG) index and the MapReduce programming framework is developed to improve the efficiency of the data processing. Chapter 15 analyses a life-log comprising image data and context and applies algorithms to collate, mine and categorise the data. Four approaches were investigated and applied to each participant's data-set, yielding an average 84.02% reduction in size. Chapter 16 proposes a privacy-enabled smart home framework consists of three major components: privacy-preserving data management module, activity recognition and occupancy detection and voice assistant. The chapter also

presents a detailed description of system architecture with service middleware of the proposed framework.

As such, this comprehensive and timely book is conceived as a unique and essential reference for the subject of smart assisted living, providing further research opportunities in this dynamic field. It is hoped that this book will provide resources necessary for policy makers, educators, students, technology developers and managers to adopt and implement smart assisted living systems. The book aims to attract four specific types of public groups interested in Smarter Assisted Living solutions: (1) end users, carers, clinicians, healthcare service providers: to inform them about smart assisted living technologies and solutions and their benefits to improve people's quality of live and independence; (2) industrial, academic and commercial organisations: to enhance the transfer of knowledge and seek for opportunities regarding commercialisation and exploitation of research outcomes; (3) government policy makers, funding bodies: to define joint RTD strategies and priorities; and (4) research student on Ambient Assisted Living, Smart Home technology, e-healthcare.

Leicester, UK  
Madrid, Spain  
Belfast/Leicester, UK  
Madrid, Spain  
Newtownabbey, UK

Feng Chen  
Rebeca I. García-Betances  
Liming Chen  
María Fernanda Cabrera-Umpiérrez  
Chris Nugent

# Contents

## Part I Sensing and Monitoring Technologies

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Pervasive Sensing</b> . . . . .  | <b>3</b>  |
|          | Yiqiang Chen  |           |
| <b>2</b> | <b>Indoor Localization and Human Activity Tracking with Multiple Kinect Sensors</b> . . . . .                     | <b>23</b> |
|          | Shunkun Yang, Akshat Hans, Wenbing Zhao and Xiong Luo   |           |
| <b>3</b> | <b>Unobtrusive Sensing Solution for Post-stroke Rehabilitation</b> . . . . .                                      | <b>43</b> |
|          | Idongesit Ekerete, Chris Nugent, Oonagh M. Giggins and James McLaughlin   |           |
| <b>4</b> | <b>Lessons from Hands-Free Data Entry in Flexible Cystoscopy with Glass for Future Smart Assistance</b> . . . . . | <b>63</b> |
|          | Charles Templeman, Francisco Javier Ordoñez Morales, Mathias Ciliberto, Andrew Symes and Daniel Roggen            |           |

## Part II Activity Recognition and Behaviour Analysis

|          |   |            |
|----------|---|------------|
| <b>5</b> | <b>Human Activity Identification in Smart Daily Environments</b> . . . . .  | <b>91</b>  |
|          | Hossein Malekmohamadi, Nontawat Pattanajak and Roeland Bom  |            |
| <b>6</b> | <b>Object Detection-Based Location and Activity Classification from Egocentric Videos: A Systematic Analysis</b> . . . . .    | <b>119</b> |
|          | Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas P. J. J. Noldus and Remco C. Veltkamp                                  |            |
| <b>7</b> | <b>Improving the Collection and Understanding the Quality of Datasets for the Aim of Human Activity Recognition</b> . . . . . | <b>147</b> |
|          | Angelica Poli, Susanna Spinsante, Chris Nugent and Ian Cleland  |            |

|  |   |     |
|--|---|-----|
| <b>8</b>   | <b>Automated General Movement Assessment for Perinatal Stroke Screening in Infants</b> .....                                    | 167 |
|  | Yan Gao, Yang Long, Yu Guan, Anna Basu, Jessica Baggaley and Thomas Plötz   |     |
| <b>Part III User Needs and Personalisation</b>             |   |     |
| <b>9</b>   | <b>User-Centered Design in Defining and Developing Health and Well-Being ICT Solutions</b> .....                                | 191 |
|  | Nikolaos Liappas, José G. Teriús-Padrón, Rebeca I. García-Betances, María Fernanda Cabrera-Umpiérrez and María Teresa Arredondo |     |
| <b>10</b>  | <b>Service Robot Behaviour Adaptation Based on User Mood, Towards Better Personalized Support of MCI Patients at Home</b> ..... | 209 |
|  | Dimitrios Giakoumis, Georgia Peleka, Manolis Vasileiadis, Ioannis Kostavelis and Dimitrios Tzovaras                             |     |
| <b>11</b>  | <b>Using Semantic Markup to Boost Context Awareness for Assistive Systems</b> .....   | 227 |
|  | Claudia Steinberger and Judith Michael  |     |
| <b>Part IV Open Smart Home and Service Infrastructures</b> |   |     |
| <b>12</b>  | <b>Towards Multi-resident Activity Monitoring with Smarter Safer Home Platform</b> .....  | 249 |
|  | Son N. Tran and Qing Zhang  |     |
| <b>13</b>  | <b>New Environments for the Evaluation of Smart Living Solutions</b> .....  | 269 |
|  | Beatriz Merino Barbancho, Ivana Lombroni, Cecilia Vera-Muñoz and María Teresa Arredondo   |     |
| <b>14</b>  | <b>A Distributed Spatial Index on Smart Medical System</b> .....  | 287 |
|  | Changqing Ji, Yang Gao, Zumin Wang and Jing Qin   |     |
| <b>15</b>  | <b>Data Reduction Methods for Life-Logged Datasets</b> .....  | 305 |
|  | William P. Burns, Paul J. McCullagh, Dewar D. Finlay, Cesar Navarro-Paredes and James McLaughlin                                |     |
| <b>16</b>  | <b>Privacy-Enabled Smart Home Framework with Voice Assistant</b> .....  | 321 |
|  | Deepika Singh, Ismini Psychoula, Erinc Merdivan, Johannes Kropf, Sten Hanke, Emanuel Sandner, Liming Chen and Andreas Holzinger |     |
|  | <b>Index</b> .....  | 341 |

**Part I**  
**Sensing and Monitoring Technologies**



# Chapter 1

## Pervasive Sensing



Yiqiang Chen

**Abstract** The development of chips, sensors, and tele-communication, etc., with integrated sensing brings more opportunities to monitor various aspects of personal behavior and context. Especially, with the widespread use of intelligent devices and smart home infrastructure, it is more possible and convenient to sense users' daily life. Two common information of daily life is location and activity. Location information can reveal the places of important events. Activity information can expose users' health conditions. Besides these two kinds of information, other context also can be useful for assisting living. Hence, in this chapter, we will introduce some state-of-the-art user context sensing techniques under smart home infrastructure, including accurate indoor localization, fine-grained activity recognition, and pervasive context sensing. With the continuous sensing of location, activity, and other contextual information, it is possible to discovery users' life patterns which are crucial for health monitoring, therapy, and other services. What is more, it will bring more opportunities for improving the quality of peoples' life.

**Keywords** Pervasive sensing · Indoor localization · Activity recognition · Context sensing

### 1.1 Accurate Indoor Localization

Do you know how to accurately get you location information under unpredictable changes in environmental conditions? In recent years, with the development of mobile Internet, location-based services (LBSs) [1] have been widely used in our daily life, expanded from traditional navigation to real-time applications such as shared mobility and social network. With the development of LBS applications, the location area extends from outdoors to indoors, which creates great requirement of indoor localization with high accuracy. Indoor localization can be implemented in a variety of ways, such as base station, video, infrared, Bluetooth, Wi-Fi [2]. In which, Wi-Fi-based indoor localization has become the most popular way because of the wide

---

Y. Chen (✉)

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China  
e-mail: [yqchen@ict.ac.cn](mailto:yqchen@ict.ac.cn)

© Springer Nature Switzerland AG 2020

F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications and Networks, [https://doi.org/10.1007/978-3-030-25590-9\\_1](https://doi.org/10.1007/978-3-030-25590-9_1)

coverage of Wi-Fi access points and the rapid development of intelligent terminals [3–5]. Although the research of indoor localization based on Wi-Fi has made great progress, in highly dynamic environments, due to the influence of multipath effect, environment changing and personnel flows, the fluctuation of wireless signal is large. High accuracy indoor localization still faces the problems of (1) the lack of large-scale labeled data in data layer, (2) the fluctuation of signal strength in feature layer, and (3) the weak adaption ability in model layer, which resulting in low location accuracy, rough trajectory granularity, and weak robustness. For the challenges above, this section will introduce some accurate indoor localization techniques.

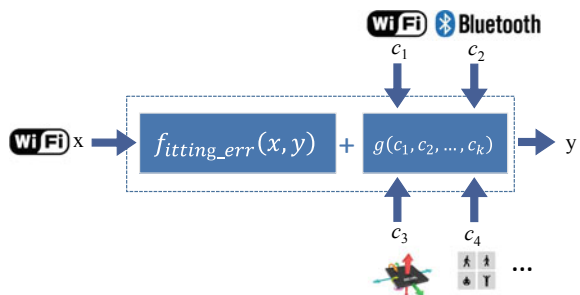
### 1.1.1 Context-Adaptive Localization Model

The wireless signal fingerprint-based indoor localization model is actually a mapping between the high-dimensional signal space and the physical space. For this kind of mapping models, the input  $\mathbf{x}$  is the feature vector extracted from the wireless signal strength, and the output  $\mathbf{y}$  is the position coordinate. Training the location model is to optimize the objective function  $f = \operatorname{argmin}_f \sum_{i=1}^N |f(x_i) - y_i|^2$  with the given samples  $\{(x_i, y_i) | i = 1, \dots, N\}$ .

However, for highly dynamic environments, a context-adaptive model is necessary. This adaptive model should include the minimization of fitting errors and the context-adaptive constraints, as shown in Fig. 1.1, where  $f_{\text{fitting\_err}}(x, y)$  represents the fitting errors between model's output and calibration results, and  $g(c_1, c_2, \dots, c_k)$  represents the constraints constructed with multi-source information of  $c_1, c_2, \dots, c_k$ . In addition, it is flexible to construct these constraints' information according to specific scenarios context, including multi-source signals, motion information, and user activities.

Compared with existing methods, the model has three advantages: (1) It gives a unified optimization objective, providing a reference for constructing multi-source information fusion localization method; (2) it realizes multi-source information fusion on the model level, fully mining the correlation and redundancy between

**Fig. 1.1** Context-adaptive location model for high dynamic environment



multi-source information; (3) it has more flexible constraints, making the model scalable for any kind of high dynamic environments.

### 1.1.2 *Semi-supervised Localization Model with Signals' Fusion*

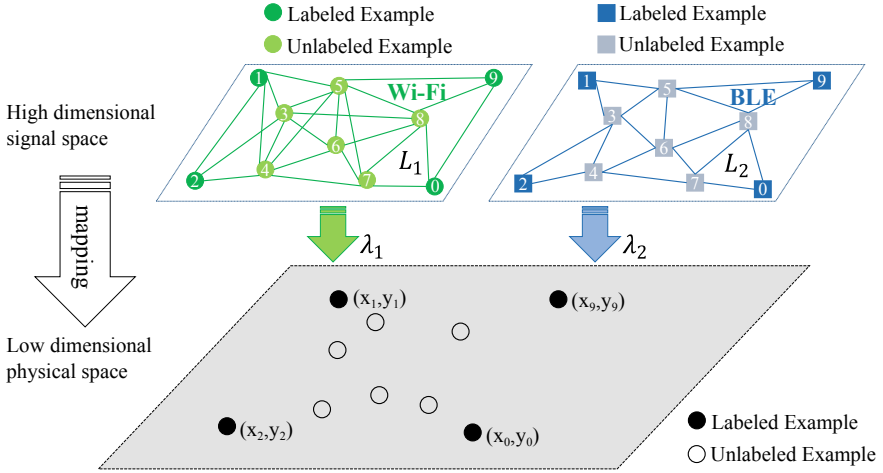
Aiming at the problem of low location accuracy caused by the lack of large-scale labeled data, a semi-supervised localization model based on multi-source signals fusion is introduced here. This model combines the fitting error term of the labeled data and manifold constraint terms of the Wi-Fi and Bluetooth signals and optimizes the objective equation by adjusting the weight coefficient of all terms. The experimental results [6] showed that the method based on multi-source signals fusion can achieve optimal location results when applied to the location problem of sparse calibration, and the location accuracy was higher than that of the existing supervised learning methods and semi-supervised learning methods.

Unlike previous single-signal-based semi-supervised manifold methods [7–12], it is better to combine the Wi-Fi and BLE signals into a single model. To the best of our knowledge, Wi-Fi and BLE signals have different propagation characteristics and effective distances. When considering both of Wi-Fi and BLE in a semi-supervised learning model, it should separately build the manifold regularization for each of them.

In accordance with the structural risk minimization principle [13], FSELM [6] used graph Laplacian regularization to find the structural relationships of both the labeled and unlabeled samples in the high-dimensional signal space. For the construction of a semi-labeled graph  $G$  based on  $l$  labeled samples and  $u$  unlabeled samples, each collected signal vector  $s_j = [s_{j1}, s_{j2}, \dots, s_{jN}] \in R^N$  is represented by a vertex  $j$ , and if the vertex  $j$  is one of the neighbors of  $i$ , represented by drawing an edge with a weight of  $w_{ij}$  connecting them. According to Belkin et al. [14], the graph Laplacian  $L$  can be expressed as  $L = D - W$ . Here,  $W = [w_{ij}]_{(l+u) \times (l+u)}$  is the weight matrix, where  $w_{ij} = \exp(-\|s_i - s_j\|^2 / 2\sigma^2)$  if  $s_i$  and  $s_j$  are neighbors along the manifold and  $w_{ij} = 0$  otherwise, and  $D$  is a diagonal matrix given by  $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$ . As illustrated in Fig. 1.2, to consider the empirical risk while controlling the complexity, FSELM minimized the fitting error plus two separate smoothness penalties for Wi-Fi and BLE as (1.1):

$$\operatorname{argmin}_f \left\{ \frac{1}{2} \|\mathbf{f} - \mathbf{T}\|^2 + \lambda_1 \mathbf{f}^T \mathbf{L}_1 \mathbf{f} + \lambda_2 \mathbf{f}^T \mathbf{L}_2 \mathbf{f} \right\} \quad (1.1)$$

The first term represents the empirical error with respect to the labeled training samples. The second and third terms represent the manifold constraints for Wi-Fi and BLE based on the graph Laplacians  $L_1$  and  $L_2$ . By adjusting the two coefficients



**Fig. 1.2** FSELM model illustration ( $L_1$  and  $L_2$  are the graph Laplacians of the Wi-Fi and BLE signals, and  $\lambda_1$  and  $\lambda_2$  are the weight coefficients of the two manifold constraints)

$\lambda_1$  and  $\lambda_2$ , it can control the relative influences of the Wi-Fi and BLE signals on the model.

When applied to sparsely calibrated localization problems, FSELM is advantageous in three aspects. Firstly, it dramatically reduces the human calibration effort required when using a semi-supervised learning framework. Secondly, it uses fused Wi-Fi and BLE fingerprints to markedly improve the location accuracy. Thirdly, it inherits the beneficial properties of ELMs in terms of training and testing speed because the input weights and biases of hidden nodes can be generated randomly. The findings indicate that effective multi-data fusion can be achieved not only through data layer fusion, feature layer fusion, and decision layer fusion but also through the fusion of constraints within a model. In addition, for semi-supervised learning problems, it is necessary to combine the advantages of different types of data by optimizing the model's parameters. These two contributions will be valuable for solving other similar problems in the future.

### 1.1.3 Motion Information Constrained Localization Model

For Wi-Fi fingerprint-based indoor localization, the basic approach is to fingerprint locations of interest with vectors of RSS of the access points during offline phase and then locate mobile devices by matching the observed RSS readings against this database during online phase. By this way, continuous localization can be summarized as trying to find a smooth trajectory going through all labeled points. Thus, in order to recover the trajectory, it still needs a certain number of labeled data, especially in some important positions (e.g., corners).

Considering that a user holds a mobile phone and walks in an indoor wireless environment with  $n$  Wi-Fi access points inside. At some time  $t$ , the signal received from all the  $n$  access points is measured by the mobile devices to form a signal vector  $s_t = [s_{t1}, s_{t2}, \dots, s_{tn}] \in R^n$ . As time goes on, the signal vectors will come in stream manner. After a period of time, a sequence of  $m$  vectors will be obtained from mobile phone and form a  $m \times n$  matrix  $S = [s_1^T, s_2^T, \dots, s_m^T]$ , where the ‘T’ indicates matrix transposition. Along the user’s trajectory, only some places are known and labeled, and the rest are unknown. The purpose is to generate and update the trajectory points which can form a  $m \times 2$  matrix  $P = [p_1^T, p_2^T, \dots, p_m^T]$ , where  $p_t = [x_t, y_t]^T$  is the location of mobile device at time  $t$ . Meanwhile, for each step, the user heading orientation can also be obtained from mobile devices in every time  $t$ . Thus, while collecting the RSS, another vector of  $m$  orientation values can be generated:  $O = [o_1, \dots, o_t, \dots, o_m]^T$ . Here,  $o_t$  indicates the angle to north in horizontal plane, which is called azimuth. With the Wi-Fi signal matrix and the orientation vector, the mapping function should be  $f(S, O) = P$ . By this way, it can supplement the location for these unlabeled data, reducing the calibration work.

The fusion mapping model  $f(S, O) = P$  from the signal space to the physical space can be optimized by  $f^* = \operatorname{argmin}_f \sum_{i=1}^l |f_i - y_i|^2 + \delta \sum_{i=1}^l |o_{f_i} - o_{y_i}|^2 + \gamma f^T L f$ , where the first term measures the fitting error to the labeled points, the second term is the fitting error to the user heading orientation offered by mobile phone, and the third term refers to the manifold graph Laplacian.

It brings good performance for both tracking mobile nodes and manual calibration reduction in wireless sensor networks. This model is based on two observations: (1) similar signals from access points imply close locations; (2) both labeled data positions and the real-time orientations can help tracking the traces. Thus, it learned a mapping function between the signal space and the physical space conjoin a few labeled data and a large amount of unlabeled data, and the constraint of orientation obtained from mobile devices.

The experimental results [15] showed that this method can achieve a higher tracking accuracy with much less calibration effort. It is robust to reduce the number of calibrated data. Furthermore, if applying it for offline calibration, the online location is much better than some other methods before. Moreover, it can reduce time consumption by parallel processing while maintaining trajectory learning accuracy.

## 1.2 Fine-Grained Activity Recognition

Traditional activity recognition methods aim at discovering pre-defined activity with body-attached sensors such as accelerometers and gyroscopes. However, peoples’ activities are so diverse; they cannot be covered by some pre-defined activities. As the way the devices are worn, the location the devices are placed, the person who wears the devices, etc., which all lead to the decreasing the recognition accuracy. And it needs a large amount of labeled data to maintain the recognition performance.

In this section, we will show the methods including transfer learning, generative adversarial networks (GANs), incremental learning to implement fine-grained activity recognition with less human labor.

### ***1.2.1 Transfer Learning-Based Activity Recognition***

The combination of sensor signals from different body positions can be used to reflect meaningful knowledge such as a person's detailed health conditions [16] and working states [17]. However, it is nontrivial to design wearing styles for a wearable device. On the one hand, it is not comfortable to equip all the body positions with sensors which make the activities restricted. Therefore, we can only attach sensors on limited body positions. On the other hand, it is impossible to perform HAR if the labels on some body parts are missing, since the activity patterns on specific body positions are significant to capture certain information.

Assume a person is suffering from small vessel disease (SVD) [18], which is a severe brain disease heavily related to activities. However, we cannot equip his all body with sensors to acquire the labels since this will make his activities unnatural. We can only label the activities on certain body parts in reality. If the doctor wants to see his activity information on the arm (we call it the target domain), which only contains sensor readings instead of labels, how to utilize the information on other parts (such as torso or leg, we call them the source domains) to help obtain the labels on the target domain? This is referred to as the cross-position activity recognition (CPAR).

To tackle the above challenge, several transfer learning methods have been proposed [19]. The key is to learn and reduce the distribution divergence (distance) between two domains. With the distance, we can perform source domain selection as well as knowledge transfer. Based on this principle, existing methods can be summarized into two categories: exploiting the correlations between features [20, 21], or transforming both the source and the target domains into a new shared feature space [22–24].

Existing approaches tend to reduce the global distance by projecting all samples in both domains into a single subspace. However, they fail to consider the local property within classes [25]. The global distance may result in loss of domain local property such as the source label information and the similarities within the same class. Therefore, it will generate a negative impact on the source selection as well as the transfer learning process. It is necessary to exploit the local property of classes to overcome the limitation of global distance learning.

This chapter introduces a Stratified Transfer Learning (STL) framework [26] to tackle the challenges of both source domain selection and knowledge transfer in CPAR. The term 'stratified' comes from the notion of splitting at different levels and then combining. The well-established assumption that data samples within the same class should lay on the same subspace, even if they come from different domains [27] is adopted. Thus, 'stratified' refers to the procedure of transforming features

into distinct subspaces. This has motivated the concept of stratified distance (SD) in comparison to traditional global distance (GD). STL has four steps:

1. **Majority Voting:** STL uses the majority voting technique to exploit the knowledge from the crowd [28]. The idea is that one certain classifier may be less reliable, so we assemble several different classifiers to obtain more reliable pseudo labels. To this end, STL makes use of some base classifiers learned from the source domain to collaboratively learn the labels for the target domain.
2. **Intra-class Transfer:** In this step, STL exploits the local property of domains to further transform each class of the source and target domains into the same subspace. Since the properties within each class are more similar, the intra-class transfer technique will guarantee that the transformed domains have the minimal distance. Initially, source domain and target domain are divided into  $C$  groups according to their (pseudo) labels, where  $C$  is the total number of classes. Then, feature transformation is performed within each class of both domains. Finally, the results of distinct subspaces are merged.
3. **Stratified Domain Selection:** A greedy technique is adopted in STL-SDS. We know that the most similar body part to the target is the one with the most similar structure and body functions. Therefore, STL uses the distance to reflect their similarity. It calculates the stratified distance between each source domain and the target domain and selects the one with the minimal distance.
4. **Stratified Activity Transfer:** After source domain selection, the most similar body part to the target domain can be obtained. The next step is to design an accurate transfer learning algorithm to perform activity transfer. This chapter introduces a Stratified Activity Transfer (STL-SAT) method for activity recognition. STL-SAT is also based on our stratified distance, and it can simultaneously transform the individual classes of the source and target domains into the same subspaces by exploiting the local property of domains. After feature learning, STL can learn the labels for the candidates. Finally, STL-SAT will perform a second annotation to obtain the labels for the residuals.

## 1.2.2 *GAN-Based Activity Recognition*

Transfer learning methods are effective ways to label practical unknown data, but they are incapable of generating realistic data. But fortunately, GANs framework is an effective way to generate labeled data from random noise space.

The vanilla GANs framework was firstly proposed in 2014 by Goodfellow et al. [29]. Since the GANs framework was proposed, it has been widely researched in many fields, such as image generation [29], image inpainting [30], image translation [31], super-resolution [32], image de-occlusion [33], natural language generation [34], text generation [35]. In particular, a great many variants of GANs have been widely explored to generate images with high fidelity, such as NVIDIA's progressive GAN [36], Google Deep Mind' BigGAN [37]. These variants of GANs provide

powerful methods for training resultful generative models that could output very convincing verisimilar images.

The original GANs framework is composed by a generative multilayer perceptron network and a corresponding discriminative multilayer perceptron network. The final goal of GANs is to estimate an optimal generator that can capture the distribution of real data with the adversarial assistance of a paired discriminator based on min-max game theory. The discriminator is optimized to differentiate the data distribution between authentic samples and spurious samples from its mutualistic generator. The generator and the discriminator are trained adversarially to achieve their optimization.

The optimization problem of the generator can be achieved by solving the formulation stated in 1.2:

$$\min_G V_G(D, G) = \min_G (\mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1.2)$$

The optimization problem of the discriminator can be achieved by solving the formulation stated in 1.3:

$$\max_D V_D(D, G) = \max_D (\mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1.3)$$

The final value function of the min-max game between the generator and the discriminator can be formulated as 1.4:

$$\min_G \max_D (D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1.4)$$

Firstly, the original GANs framework was proposed to generate plausible fake images approximating real images in low resolution, such as MNIST, TFD, CIFAR-10. Many straightforward extensions of GANs have demonstrated and leded one of the most potential research directions. Though the researches of GANs have gained great success in the field of generating realistic-looking images, the GANs framework has not been widely exploited for generating sensor data.

Inspired by the thought of GANs, Alzantot et al. [38] firstly tried idea of GANs to train the LSTM-based generator to produce sensor data, but their SenseGen is half-baked GANs' framework. Both the generator and the discriminator in SenseGen are trained separately; that is, the training process of the generator in SenseGen is not based on the back-propagating gradient from the discriminator.

In order to improve the performance of human activity recognition when a small number of sensor data are available under some special practical scenarios and resource-limited environments, Wang et al. [39] proposed SensoryGANs models. To the best of our knowledge, SensoryGANs models are the first unbroken generative adversarial networks applied in generating sensor data in the HAR research field. The specific GANs models were designed for three human daily activities, respectively. The generators accept the Gaussian random noises and output accelerometer data of the target human activity. The discriminators accept both the real accelerometer



sensor data and the spurious accelerometer sensor data from the generators and then output the probability of whether the input samples are from the real distribution. With the improvement of SensoryGANs, the research of human activity recognition, especially in resource-constrained environments, will be greatly encouraged.

Then, Yao et al. [40] proposed SenseGAN to leverage the abundant unlabeled sensing data, thereby minimizing the need for labeling effort. SenseGAN jointly trains three components, the generator, the discriminator, and a classifier. The adversarial game among the three modules can achieve their optimal performance. The generator receives random noises and labels and then outputs spurious sensing data. The classifier accepts sensing data and outputs labels. The samples from the classifier and the generator are both fed to the discriminator for differentiating the joint data/label distribution between real sensing data and spurious sensing data. Compared with supervised counterparts as well as other supervised and semi-supervised baselines, SenseGAN achieves substantial improvements in accuracy and F1 score. With only 10% of the originally labeled data, SenseGAN can attain nearly the same accuracy as a deep learning classifier trained on the fully labeled dataset.

### ***1.2.3 Incremental Learning-Based Activity Recognition***

With more labeled data, it becomes possible to get fine-grained activity. However, traditional sensor-based activity recognition methods train fixed classification models with labeled data collected off-line, which are unable to adapt to dynamic changes in real applications. With the emergence of new wearable devices, more diverse sensors can be used to improve the performance of activity recognition. While it is difficult to integrate a new sensor into a pre-trained activity recognition model, the emergence of new sensors will lead to a corresponding increase in the feature dimensionality of the input data, which may result in the failure of a pre-trained activity recognition model. The pre-trained activity recognition model is unable to take advantage of this new source of data.

To take advantage of data generated by new sensors, feature incremental learning method is an effective method. To improve the performance of indoor localization with more sensors, Jiang et al. [41] proposed a novel feature incremental and decremental learning method, namely FA-OSELM. It is able to adapt to the dynamic changes of sensors flexibly. However, the performance of FA-OSELM fluctuates heavily. Hou and Zhou [42] proposed the One-Pass Incremental and Decremental learning approach (OPID), which is able to adapt to evolving features and instances simultaneously. Xing et al. [43] proposed a perception evolution network that integrates the new sensor readings into the learned model. However, the impact of the sensor order is not considered.

Hu et al. [44] proposed a novel feature incremental activity recognition method, which is named Feature Incremental Random Forest (FIRF). It is able to adapt an existing activity recognition model to newly available sensors in a dynamic environment. Figure 1.3 shows an overview of the method.

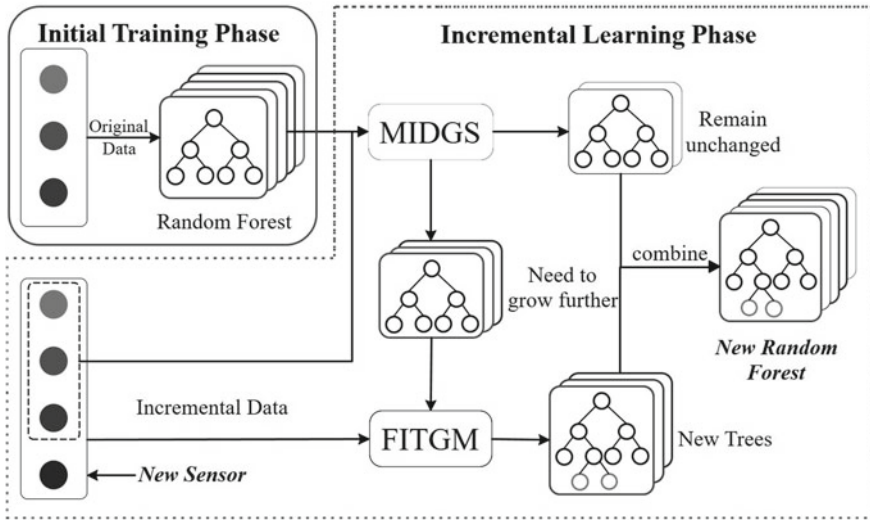


Fig. 1.3 Feature incremental random forest algorithm

In FIRE, there are two new strategies: (1) MIDGS which encourages diversity among individual decision trees in the incremental learning phase by identifying the individual learners that have high redundancy with the other individual learners and low recognition accuracy, and (2) FITGM which improve the performance of these identified individual decision trees with new data collected from both existing and newly emerging sensors.

In real applications, people may learn new motion activities over time, which is usually classified as dynamic changes in class. When a new kind of activity is performed or the behavioral pattern changes over time, devices with preinstalled activity recognition models may fail to recognize new activities or even known activities with changed manners. To adapt to the changes of activities, traditional batch learning methods require retraining the whole model from scratch. This will result in a great waste of time and memory.

Class incremental learning method is an effective way to address this problem. Different from batch learning, incremental learning, or online learning methods update existing models with new knowledge. In [45], Zhao et al. presented a class incremental extreme learning machine (CIELM), which adds new output nodes to accommodate new class data. With update to output weights, CIELM can recognize new activities dynamically. Camoriano et al. [46] employed recursive technique and regularized least squares for classification (RLSC) to seamlessly add new classes to the learned model. They considered the imbalance between classes in the class incremental learning phase. Zhu et al. [47] introduced a framework, namely the one-pass class incremental learning (OPCIL), to handle new emerging classes. They proposed a pseudo instances generating approach to address the new class adaptation issue. Ristin et al. [48] put forward two variants of random forest to incorporate new classes.

Four incremental learning strategies are devised to exploit hierarchical nature of random forest for efficient updating.

In [49], Hu et al. proposed an effective class incremental learning method, named class incremental random forest (CIRF), to enable existing activity recognition models to identify new activities. They designed a separating axis theorem-based splitting strategy to insert internal nodes and adopt Gini index or information gain to split leaves of the decision tree in the random forests. With these two strategies, both inserting new nodes and splitting leaves are allowed in the incremental learning phase. They evaluated their method on three UCI public activity datasets and compared with other state-of-the-art methods. Experimental results show that their incremental learning method converges to the performance of batch learning methods (random forests and extremely randomized trees). Compared with other state-of-the-art methods, it is able to recognize new class data continuously with a better performance.

### 1.3 Pervasive Context Sensing

With the pervasiveness of intelligent hardware, more individual context can be sensed, which is meaningful to infer users' life patterns, health conditions, etc. In this section, we will introduce context sensing methods with pervasive intelligent hardware, including sleep sensing, household water-usage sensing, etc.

#### 1.3.1 *Sleeping Sensing*

Sleeping is a vital activity that people spend nearly a third of lifetime to do. Many studies have shown that sleep disorder is related to many serious diseases including senile dementia, obesity, and cardiovascular disease [50]. Clinical studies have reported that sleeping is composed of two stages including rapid eye movement (REM) and non-rapid eye movements (NREM). NREM can be further divided into light and deep sleep stages. During sleep, REM and NREM change alternately. Babies can spend up to 50% of their sleep in the REM stage, compared to only about 20% for adults. As people getting older, they sleep more lightly and get less deep sleep. Therefore, it is meaningful to find out the distribution of different sleep stages.

As sleep quality is very important for health, a lot of previous researches have been done on sleep detection. The methods of analyzing sleep quality mainly monitor different sleep stages. Recently, the technologies of recording sleep stages are divided into two categories. One category is polysomnography (PSG)-based approaches [51]. PSG monitors many body functions including brain (EEG), eye movements (EOG), skeletal muscle activation (EMG), and heart rhythm (ECG) during sleep. However, collecting the polysomnography signals or brain waves requires professional equipments and specialized knowledge. Another category is actigraphy-based approaches. Typical devices can be divided into the following two categories. The first category

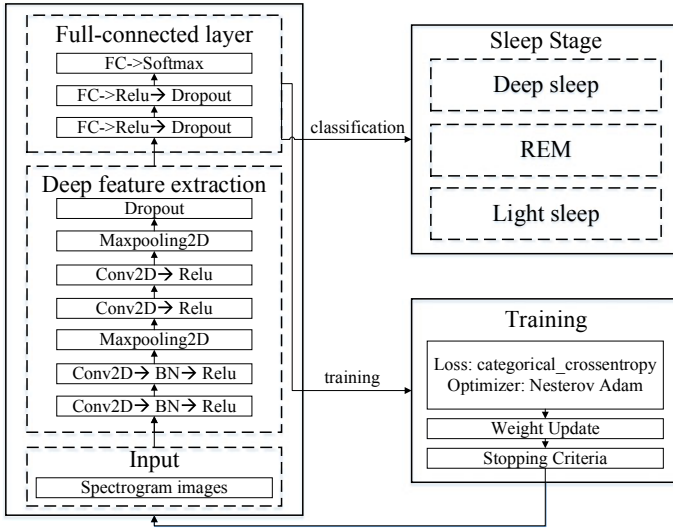
is wearable sleep and fitness tracker such as Fitbit charge 2 and Jawbone Up [52]. These devices primarily work by actigraphy. Several algorithms [53] utilized wrist activity data to predict sleep/wake states. The results have shown that the accuracy of predicting sleep/wake through recording wrist activity data approaches score using EEG data. But wearable sleep devices have some weaknesses because of accuracy concerns for sleep stages. These devices detect sleep stages based on logged acceleration data generated by body movement. This means if a user does not move, these devices have to rely on other auxiliary sensors. The second category is non-wearable sleep trackers such as Beddit 3.0 Smart Sleep Monitor. These are dedicated sleep trackers that users do not wear on wrist. They tend to provide more detailed sleep data. Many products use non-wearable smartphone sensors to assess sleep quality or sleep stage. An application called iSleep [54] used the microphone of smartphone to detect the sleep events. The method extracts three features to classify different events including body movement, snoring, and coughing. These non-wearable sleep trackers tend to use many sensors on smartphone and a lot of manual intervention to extract features.

Different from these works, the work [55] leveraged microphone without any other auxiliary sensor or much manual intervention to detect sleep stages. Acoustic signal collected by the microphone is sensitive enough to record information. After the acoustic signal is collected, the spectrogram visual representation is given. Specifically, the spectrogram is the magnitude squared of the short-time Fourier transform (STFT). It splits time signal to short segments of equal length and then computes STFT on each segment.

Once the spectrogram has been computed, they can be processed by the deep learning model. Deep learning is a new area of machine learning research. Its algorithms build a large set of layers to extract a hierarchy of features from low level to high level. Deep learning models include deep neural network (DNN), convolutional neural network (CNN, or ConvNet), etc. ConvNet [56] is the most efficient approach for image and speech recognition. The major difference between ConvNet and ordinary neural networks is that ConvNet architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture and vastly reduce the amount of parameters in the network.

The convolutional neural network architecture and training procedure are shown in Fig. 1.4. Learning from the relatively good effect of the network configuration in image recognition, this configuration can improve the expression ability of ConvNet. At the same time, accumulating convolutional layers and pooling layers guarantees long-range dependence (LRD) of acoustic signal, which is more robust than conventional ConvNet architecture.

During the training process, the goal is to minimize the loss function in backward propagation. The optimizers such as stochastic gradient descent (SGD) and Nadam are used to update the weights of hidden layers. The output of network is divided into three categories, deep sleep, light sleep, and REM.



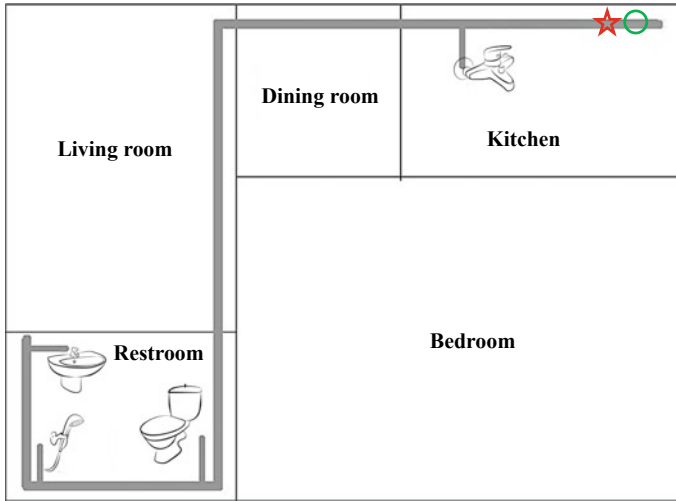
**Fig. 1.4** Illustration of the convolutional neural networks (CNNs) architecture and training procedure

### 1.3.2 Water-Usage Sensing

A person’s daily activities can be recognized by monitoring the infrastructures (e.g., water, electric, heat, ventilating, air conditioning) in the house. Infrastructure-mediated sensing has been recognized as a low-cost and nonintrusive activity recognition technique.

Several infrastructure-mediated sensing approaches for water-usage activity recognition have been proposed recently. A water-usage activity recognition technique was proposed by Fogarty [57], deploying four microphones on the surface of water pipes near the inlets and outlets. Froehlich et al. [58] proposed HydroSense, another infrastructure-mediated single-point sensing technique. Thomaz et al. [59] proposed an learning approach for high-level activity recognition, which combined single-point, infrastructure-mediated sensing with a vector space model. Their work has been considered to be the first one of employing the method for inferring high-level water-usage activities. However, the infrastructure of the house has to be remodeled in order to work in with the installation of the pressure sensors.

To solve the above question, a single-point infrastructure-mediated sensing system for water-usage activity recognition proved to be effective which has a single 3-axis accelerometer clinging to the surface of the main water pipe in the house [60]. The structure of the water pipe in the apartment can be seen in Fig. 1.5. The thick and thin gray lines represent the main water pipe and branches of the main pipe. The green circle and red star in Fig. 1.5 are water meter and the accelerometer, respectively.



**Fig. 1.5** Structure of the water pipe

The water-usage activity recognition system has six modules, which are:

**A. Data Preprocessing**

Normally, there exist some noises in the raw time series samples which should be filtered out. The median filter technique is employed in this data preprocessing module, and the filter window is set to 3.

**B. Segmentation**

The segmentation module is aimed at segmenting both the rugged segments (time series rugged samples) and smooth segments (time series smooth samples) from time series samples.

First, sample windows are generated on the set of time series samples according to the sliding window mechanism; second, annotate each sample window to be rugged or smooth based on whether its standard deviation is no less than a threshold or not. At last, a rugged (or smooth) segment is defined as a time series rugged (or smooth) windows.

**C. Data Post-processing**

The data post-processing module is to make all the rugged segments generated in the previous module completer and more precise.

First-stage post-process procedure: In the first circumstance, any smooth segment (in between two rugged segments), whose corresponding samples are no more than a threshold, is re-annotated to be rugged segments. After that, all the neighboring rugged segments make up a long-rugged segment.

Second-stage post-process procedure: In the second one, any rugged segment (in between two smooth segments), whose corresponding samples are no more than another threshold, is re-annotated to be smooth segments. After that, all the neighboring smooth segments make up a long smooth segment.

#### D. *Feature Extraction*

Instances are generated by utilizing the sliding window mechanism again on each rugged segment. The feature extraction module is executed on each sub-segment. Eight features (0.25-quantile, 0.5-quantile, 0.75-quantile, mean value, standard deviation, quadratic sum, zero-crossing, spectral peak) are extracted from a window of sample values in each axis ( $x$ -axis,  $y$ -axis, or  $z$ -axis in the accelerometer device). In all, there are 24 features for each instance.

#### E. *Model Generation and Prediction*

All the instances are split into two sets (the training set and the testing set) with approximately the same size. Instances in the same segment are assured to put into the same set, since you do not want any water-usage activity to be apart.

Support vector machine (SVM) is employed for model generation, and Gaussian kernel can be utilized as its kernel function. Two parameters—the kernel parameter and the penalty parameter—need to be set before starting the learning process. In the end, a classifier is constructed on the training set.

The classifier is then employed to predict the labels of instances in the testing set (testing instances). These prediction results are recognized as SVM's prediction labels for the testing instances.

#### F. *Prediction Results' Fusion*

The prediction results' fusion module is done by law of 'The minority is subordinate to the majority'. Specifically, for each water-usage activity, the number of testing corresponds to the most testing instances. In the end, the prediction labels of all instances in the segment are replaced by the dominant water-usage activity. The prediction results of the rugged segment are fused finally.

The nonintrusive and single-point infrastructure-mediated sensing approach in this chapter can recognize 4-class water-usage activities in daily life. Data is collected unobtrusively by a single low-cost 3-axis accelerometer attached to the surface of the main water pipe in the house, making the installation process much more convenient.

### 1.3.3 *Non-contact Physiological Signal Sensing*

Non-contact vital sign detection has received significant attention from healthcare researchers, for it can perform basic physiological signal acquisition without any interference to the user. The electrode-attached ones, such as electrocardiography (ECG) or respiratory detection instrument, need the user fixed in a particular place or to be worn by the user all day long. These approaches have a negative impact on the daily life of users, which cannot be used in many applications, such as sleep apnea monitoring, burned patients' vital sign monitoring, and health care jobs that require long-term monitoring.

The heartbeats and respiratory are common physiological signals, which can be used for sleep monitoring and abnormal body monitoring. At present, the traditional

methods for heartbeat detection are electrocardiogram (ECG) and photoplethysmography (PPG). The traditional detection method for breathing is mainly measuring the air volume and flow of the nose and mouth through the breathing process. All these methods require direct physical contact with the user, and the electrodes, sensors, masks need to be placed close to the skin for physiological signal measurement. Although the measurement result is more accurate, it has a strong interference to the normal life of the user, greatly reduces the comfort of the user, and cannot achieve long-term monitoring of the physiological information of the user. Therefore, non-contact detection methods attract more interest recently.

The non-contact detection of heartbeat and respiratory can be achieved by many methods, such as camera [61], radar, Wi-Fi, ultrasonic [62]. The camera method is to perform heartbeat detection through face video and perform respiratory detection by using body video. The other methods mainly perform heartbeat and respiratory frequency detection by detecting chest vibration caused by respiratory and heartbeat. Among them, the radar method has better recognition effect when the user is still, for electromagnetic can penetrate the clothes or covers and most of it will be reflected when it reaches the surface of the human body.

The radar method also can be subdivided according to the principle of signal transmission and reception. The most used radar methods are Doppler radar, FMCW radar [63]. Also, there are many innovative radar methods are used in heartbeat and respiratory detection, such as UWB pulse radar [64], self-injection-locked radar [65], UWB impulse radar [66].

**The Doppler Radar:** The Doppler radar method measures a user's chest movement via the return signal phase. Doppler radar transmits continue wave (CW) electromagnetic wave toward the user's body, and the RF signal will be reflected from the skin and tissue of the body. The receiver acquires the electromagnetic signal and mixes the received signal with the transmitter signal for vital signal detection.

Recently, the coherent receiver is used by mixing the received signal with a quadrature mixer. The quadrature mixer mixes the original received signal and a 90-degree shifted signal with the transmitter signal to achieve two quadrature components. With this method, the NULL point of radar detection is avoided. The signal needs to be demodulation with linear or nonlinear demodulation methods to get the phase change containing  $x(t)$ . Then the heartbeat and the respiratory signal can be achieved with signal processing methods or machine learning methods.

**The FMCW Radar:** The frequency modulated continuous wave (FMCW) radar can determine the absolute distance between the system and a target. The FMCW radar transmits variable frequency signal with a modulation frequency being able to slew up and down as sine wave, sawtooth wave, triangle wave, or square wave [63]. And for vital sign detection, if the target is a man, the received signal will contain the information of the chest movement. Then, the signal will have a frequency shift of the chest motion frequency. By detecting the frequency shift in the range information, the heartbeat and respiratory frequency can be calculated.



## 1.4 Conclusions

In this chapter, we show different ways to sense user's location information, activity information, and other context information with the pervasiveness of intelligent devices under smart home infrastructure. In the future, with the development of Internet of things (IoT), edge computing, and cloud computing, the sensing ability in smart home will be unprecedentedly powerful. And the collaborative computing framework of the above three (IoT, edge computing and cloud computing) would be the trend, which can adaptively use the device and resource to optimally achieve the task, what is more, with the maturity of pervasive sensing techniques, it will bring more convenience to people's daily life and make high-quality living possible.

## References

1. Prasad M (2002) Location based services. *GIS Dev* 3–35
2. Xiang Z, Song S, Chen J et al (2004) A wireless LAN-based indoor positioning technology. *IBM J Res Dev* 48(5.6):617–626
3. Liu H, Darabi H, Banerjee P, Liu J (2007) Survey of wireless indoor positioning techniques and systems. *IEEE Trans Syst Man Cybern Part C* 37(6):1067–1080
4. Kjægaard MB (2007) A taxonomy for radio location fingerprinting. In: Hightower J, Schiele B, Strang T (eds) *Location- and contextawareness*, LNCS, vol 4718. Springer, Berlin, pp 139–156
5. Brunato M, Battiti R (2005) Statistical learning theory for location fingerprinting in wireless LANs. *Comput Netw* 47:825–845
6. Jiang X, Chen Y, Liu J, Gu Y, Hu L (2018) FSELM: fusion semi-supervised extreme learning machine for indoor localization with Wi-Fi and Bluetooth fingerprints. *Soft Comput* 22(11):3621–3635
7. Liu J, Chen Y, Liu M, Zhao Z (2011) Selm: semi-supervised elm with application in sparse calibrated location estimation. *Neurocomputing* 74(16):2566–2572
8. Pan JJ, Yang Q, Chang H, Yeung D-Y (2006) A manifold regularization approach to calibration reduction for sensor-network based tracking. In: *AAAI*, pp 988–993
9. Pan JJ, Yang Q, Pan SJ (2007) Online co-localization in indoor wireless networks by dimension reduction. In: *Proceedings of the national conference on artificial intelligence*, vol 22. Menlo Park, CA; Cambridge, MA; London; AAI Press; MIT Press; 1999, p 1102
10. Gu Y, Chen Y, Liu J, Jiang X (2015) Semi-supervised deep extreme learning machine for Wi-Fi based localization. *Neurocomputing* 166:282–293
11. Zhang Y, Zhi X (2010) Indoor positioning algorithm based on semisupervised learning. *Comput Eng* 36(17):277–279
12. Scardapane S, Comminiello D, Scarpiniti M, Uncini A (2016) A semisupervised random vector functional-link network based on the transductive framework. *Inf Sci* 364:156–166
13. Vapnik V (2013) *The nature of statistical learning theory*. Springer science & business media, Berlin
14. Belkin M, Niyogi P, Sindhvani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J Mach Learn Res* 7:2399–2434
15. Jiang X, Chen Y, Liu J, Gu Y, Hu L, Shen Z (2016) Heterogeneous data driven manifold regularization model for fingerprint calibration reduction. In: *2016 International IEEE conferences on ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, internet of people, and smart world congress (UIC/ATC/ScalCom/CBDCCom/IoP/SmartWorld)*. IEEE, pp 74–81

16. Hammerla N Y, Fisher J, Andras P et al (2015) PD disease state assessment in naturalistic environments using deep learning. In: Twenty-Ninth AAAI conference on artificial intelligence
17. Plötz T, Hammerla NY, Olivier PL (2011) Feature learning for activity recognition in ubiquitous computing. In: Twenty-second international joint conference on artificial intelligence
18. Wardlaw J M, Smith EE, Biessels GJ et al (2013) Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurol* 12(8):822–838
19. Cook D, Feuz KD, Krishnan NC (2013) Transfer learning for activity recognition: a survey. *Knowl Inf Syst* 36(3):537–556
20. Blitzer J, McDonald R, Pereira F (2006) Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 120–128
21. Kouw WM, Van Der Maaten LJP, Krijthe JH et al (2016) Feature-level domain adaptation. *J Mach Learn Res* 17(1):5943–5974
22. Pan SJ, Tsang IW, Kwok JT et al (2011) Domain adaptation via transfer component analysis. *IEEE Trans Neural Netw* 22(2):199–210
23. Gong B, Shi Y, Sha F et al (2012) Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 2066–2073
24. Long M, Wang J, Sun J et al (2015) Domain invariant transfer kernel learning. *IEEE Trans Knowl Data Eng* 27(6):1519–1532
25. Lin Y, Chen J, Cao Y et al (2017) Cross-domain recognition by identifying joint subspaces of source domain and target domain. *IEEE Trans Cybern* 47(4):1090–1101
26. Wang J, Chen Y, Hu L et al (2018) Stratified transfer learning for cross-domain activity recognition. In: 2018 IEEE international conference on pervasive computing and communications (PerCom). IEEE, pp 1–10
27. Elhamifar E, Vidal R (2013) Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans Pattern Anal Mach Intell* 35(11):2765–2781
28. Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532
29. Goodfellow I, Pouget-Abadie J, Mirza M et al (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
30. Yeh RA, Chen C, Yian Lim T et al (2017) Semantic image inpainting with deep generative models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5485–5493
31. Isola P, Zhu JY, Zhou T et al (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
32. Ledig C, Theis L, Huszár F et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
33. Zhao F, Feng J, Zhao J et al (2018) Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Trans Image Process* 27(2):778–790
34. Press O, Bar A, Bogin B et al (2017) Language generation with recurrent generative adversarial networks without pre-training. arXiv preprint [arXiv:1706.01399](https://arxiv.org/abs/1706.01399)
35. Yu L, Zhang W, Wang J et al (2017) Seqgan: sequence generative adversarial nets with policy gradient. In: Thirty-first AAAI conference on artificial intelligence
36. Karras T, Aila T, Laine S et al (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint [arXiv:1710.10196](https://arxiv.org/abs/1710.10196)
37. Brock A, Donahue J, Simonyan K (2018) Large scale gan training for high fidelity natural image synthesis. arXiv preprint [arXiv:1809.11096](https://arxiv.org/abs/1809.11096)
38. Alzantot M, Chakraborty S, Srivastava M (2017) Sensegen: a deep learning architecture for synthetic sensor data generation. In: 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, pp 188–193

39. Wang J, Chen Y, Gu Y et al (2018) SensoryGANs: an effective generative adversarial framework for sensor-based human activity recognition. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–8
40. Yao S, Zhao Y, Shao H et al (2018) Sensegan: enabling deep learning for internet of things with a semi-supervised framework. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2(3):144
41. Jiang X, Liu J, Chen Y et al (2016) Feature adaptive online sequential extreme learning machine for lifelong indoor localization. *Neural Comput Appl* 27(1):215–225
42. Hou C, Zhou ZH (2018) One-pass learning with incremental and decremental features. *IEEE Trans Pattern Anal Mach Intell* 40(11):2776–2792
43. Xing Y, Shen F, Zhao J (2015) Perception evolution network adapting to the emergence of new sensory receptor. In: Twenty-fourth international joint conference on artificial intelligence
44. Hu C, Chen Y, Peng X et al (2018) A novel feature incremental learning method for sensor-based activity recognition. *IEEE Trans Knowl Data Eng*
45. Zhao Z, Chen Z, Chen Y et al (2014) A class incremental extreme learning machine for activity recognition. *Cognit Comput* 6(3):423–431
46. Camoriano R, Pasquale G, Ciliberto C et al (2017) Incremental robot learning of new objects with fixed update time. In: 2017 IEEE international conference on robotics and automation (ICRA). IEEE, pp 3207–3214
47. Zhu Y, Ting KM, Zhou ZH (2017) New class adaptation via instance generation in one-pass class incremental learning. In: 2017 IEEE international conference on data mining (ICDM). IEEE, pp 1207–1212
48. Ristin M, Guillaumin M, Gall J et al (2016) Incremental learning of random forests for large-scale image classification. *IEEE Trans Pattern Anal Mach Intell* 38(3):490–503
49. Hu C, Chen Y, Hu L et al (2018) A novel random forests based class incremental learning method for activity recognition. *Pattern Recognit* 78:277–290
50. Kanbay A, Buyukoglan H, Ozdogan N et al (2012) Obstructive sleep apnea syndrome is related to the progression of chronic kidney disease. *Int Urol Nephrol* 44(2):535–539
51. Berry RB, Brooks R, Gamaldo CE et al (2012) The AASM manual for the scoring of sleep and associated events. In: Rules, terminology and technical specifications. Darien, Illinois, American Academy of Sleep Medicine, p 176
52. Jawbone Up <https://jawbone.com/up>
53. Hoque E, Dickerson RF, Stankovic JA (2010) Monitoring body positions and movements during sleep using wisps. In: *Wireless health 2010*. ACM, pp 44–53
54. Hao T, Xing G, Zhou G (2013) iSleep: unobtrusive sleep quality monitoring using smartphones. In: *Proceedings of the 11th ACM conference on embedded networked sensor systems*. ACM, p 4
55. Zhang Y, Chen Y, Hu L et al (2017) An effective deep learning approach for unobtrusive sleep stage detection using microphone sensor. In: 2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI). IEEE Computer Society
56. LeCun Y, Bottou L, Bengio Y et al (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
57. Fogarty J, Au C, Hudson SE (2006) Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In: *Proceedings of the 19th annual ACM symposium on user interface software and technology*. ACM, pp 91–100
58. Froehlich JE, Larson E, Campbell T et al (2009) HydroSense: infrastructure-mediated single-point sensing of whole-home water activity. In: *Proceedings of the 11th international conference on Ubiquitous computing*. ACM, pp 235–244
59. Thomaz E, Bettadapura V, Reyes G et al (2012) Recognizing water-based activities in the home through infrastructure-mediated sensing. In: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, pp 85–94
60. Hu L, Chen Y, Wang S et al (2013) A nonintrusive and single-point infrastructure-mediated sensing approach for water-use activity recognition. In: 2013 IEEE 10th international conference on high performance computing and communications & 2013 IEEE international conference on embedded and ubiquitous computing. IEEE, pp 2120–2126

61. Qi H, Guo Z, Chen X et al (2017) Video-based human heart rate measurement using joint blind source separation. *Biomed Signal Process Control* 31:309–320
62. Min SD, Kim JK, Shin HS et al (2010) Noncontact respiration rate measurement system using an ultrasonic proximity sensor. *IEEE Sens J*, 10(11):1732–1739
63. Li C, Peng Z, Huang TY et al (2017) A review on recent progress of portable short-range noncontact microwave radar systems. *IEEE Trans Microw Theory Techn* 65(5):1692–1706
64. Hosseini SMAT, Amindavar H (2017) UWB radar signal processing in measurement of heart-beat features[C]. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1004–1007
65. Wang FK, Tang MC, Su SC et al (2016) Wrist pulse rate monitor using self-injection-locked radar technology. *Biosensors* 6(4):54
66. Cho HS, Park YJ, Lyu HK et al (2017) Novel heart rate detection method using UWB impulse radar. *J Signal Process Syst* 87(2):229–239

# Chapter 2

## Indoor Localization and Human Activity Tracking with Multiple Kinect Sensors



Shunkun Yang, Akshat Hans, Wenbing Zhao and Xiong Luo

**Abstract** In this chapter, we describe a system that provides continuous localization and behavioral analysis of a person's motion pattern over an indoor living space using multiple Kinect sensors. The skeleton data from all sensors is transferred to the host computer via TCP sockets into a program where the data is integrated into a single world coordinate system using a calibration technique. Multiple cameras are placed with some overlap in the field of view for the successful calibration of the cameras and continuous tracking of the patients. Localization and behavioral data is stored in a CSV file for further analysis. The experiments show that the system can reliably detect sitting and standing poses, as well as basic gait parameters of a user who is walking within the field of view. This system may be used in an assistive living environment to track the activities of daily living of seniors.

**Keywords** Indoor localization · Human activity tracking · Microsoft Kinect · Activities of daily living · Gait

### 2.1 Introduction

The monitoring of activities of daily living of seniors in assistive living environments is crucial to protect their well-being and could potentially help us understand the progression of diseases that impact their motor and/or cognitive skills, such as mild cognitive impairment, Alzheimer's dementia, Parkinson's diseases [14, 16, 17, 42]. Understanding a person's activity is a challenging task because it requires the accurate determination of both the action (such as sitting, walking) and the context where the action is taking place (such as the dinner table in the kitchen or the couch in the living

---

S. Yang  
Beihang University, Beijing, China

A. Hans · W. Zhao (✉)  
Cleveland State University, Cleveland, USA  
e-mail: [w.zhao1@csuohio.edu](mailto:w.zhao1@csuohio.edu)

X. Luo  
University of Science and Technology Beijing, Beijing, China

room). Human action recognition has been intensely studied relative to the action context determination. One way to determine the context is via object detection, which could be very difficult due to the many objects in a residential home. Another way for doing so is by deploying beacons around a residential home, and the person who is being tracked must carry a smartphone running a beacon signal detection application. This might not be an attractive approach either due to the need of sensor deployment and the carrying of a smartphone everywhere the person goes at his or her own home.

In this chapter, we introduce a system that can determine both the action and the location where the action is taking place by using multiple Kinect sensors. We refer to the latter as indoor localization. The system contains the following major components: (1) Use multiple Kinect sensors to cover and monitor a large area; (2) record a person's trajectory along with a time stamp; (3) classify activities and determine the total time spent in an area/room and sitting/standing; and (4) provide a visual representation of classified activities. Our system uses the RoomAlive Toolkit developed by Microsoft, which provides a platform for users to calibrate and create a 3D model of any room by using multiple Kinect cameras. RoomAlive Toolkit utilizes a projector to display the gray codes for calibration of multiple Kinect sensors.

## 2.2 Related Work

### 2.2.1 Human Tracking

Tracking is the essential component for surveillance applications. Many methodologies have been proposed for tracking humans based on variations of the same Bayesian framework [20], such as Kalman filters [12], particle filters, and mean shift algorithm. Kalman filters have been utilized widely to track in many domains. A large number of works [1, 3, 22, 26, 43] have used Kalman filters for tracking. In particle filtering, the posterior is at first approximated by the arrangement of discrete samples with related weights [4]. The particle with smaller weight is disposed of in the following iteration, while those with the substantial weight are replicated to maintain the population size. The particle filter will converge on a hypothesis after several iterations. For tracking people in the cluttered scene, many papers make use of particle filtering. Breitenstein et al. [4] proposed a multi-person tracking-by-detection in a particle filtering framework. Detection and detection confidence are used for propagating the particles. The method proposed by Comaniciu et al. [7] provides basic tracking framework based on a mean shift algorithm. Tracking is done by finding the peak in the probability density function calculated on each pixel using color similarity.

Occlusion makes tracking difficult. In single camera systems, to cope with the occlusion, approaches such as the predicted position of the occluded person until the person re-appears [9] are common to find. In partial occlusion, the visible part of

the person can still be used to track the person. The camera can also be placed higher facing downward to reduce the effect of occlusion. With the continuous increase in computing power, many researchers are inclined toward the use of multiple overlapping cameras to reduce the effect of occlusion [8].

### ***2.2.2 Depth Cameras and Skeleton Tracking***

RGB-D cameras provide an additional stream of depth data along with the color stream. Human detection and tracking from color frames can be difficult due to a range of factors. Detection of humans using 2D color image can become difficult with the change in illumination, color, clutter, and occlusion as it becomes cumbersome to separate background and foreground. On the other hand, depth image greatly simplifies the problem of inconsistent color and illumination [10, 21]. Color frames represent the 3D world in a 2D image. This conversion of 3D space into the 2D image can be represented by a pinhole camera model. The depth image is a simple representation of the 3D space. Background separation is a lot easier in a depth image as compared to the 2D color image. Spinello and Arras [23] proposed people detection using a histogram of oriented depth (HOD), inspired by HOG but using depth information instead. Stereo cameras are also used for range (depth) imaging. Masuyama et al. [18] detected humans by using subtraction stereo to the images captured by a stereo camera to obtain foreground region and corresponding depth information.

During recent years, different types of depth cameras have emerged. However, earlier depth cameras were expensive. Microsoft Kinect [15, 32] was launched by Microsoft back in 2010 and was initially developed for the Xbox gaming console. Later, many computer vision research communities and other less apparent communities such as design, materials science, robotics, biology, and medicine started using the Kinect because of its reasonable depth accuracy and affordability [27–31, 33–41]. Also, the Kinect sensor provides real-time 3D human skeleton data. In skeleton tracking, the human body is rendered by a number of joints representing different body parts, every joint being represented by its 3D coordinates.

3D skeleton-based human representation also reveals promising performance in real-world applications such as Kinect-based motion controller. 3D skeleton data from Kinect is also robust to illumination changes. Skeleton detection and processing are done in the Kinect device itself to off-load the computing power required for skeleton processing from user's computer and facilitate high frame rate, real-time, online applications using the skeleton data provided by Kinect. Zhao et al. [30, 36] developed a Kinect-based rehabilitation exercise monitoring system which shows that the skeleton data provided by the Kinect can be successfully used in monitoring rehabilitation exercises.

### 2.2.3 *Tracking with Multiple Kinect Sensors*

Often in applications such as surveillance and activity tracking, large field of view is required to cover every part of the room or area being monitored. A standard solution to increase the field of view is to use camera networks along with sensor fusion. Caon et al. [5] used a simple method of skeleton data fusion, using Kinect v1 cameras for their smart environment-related application. They proposed a weighted averaging method for joint coordinate fusion. A weighting factor of the joint coordinates depends on the tracking state of the joints derived from Kinect SDK and a total number of joints tracked. In Williamson et al. [25], multi-Kinect sensor setup is used for dismounted soldier training. Cross-validation is done by comparing the joint depth ( $Z$ ) with the point cloud's depth measurement at the same  $X$ - $Y$ -coordinates. In Azis et al. [2], two perpendicular Kinect sensors are used. One sensor is defined as the primary sensor. If the primary sensor does not track a joint, coordinates of the untracked joint are substituted by the one tracked by the second Kinect sensor.

### 2.2.4 *Indoor Localization and Motion Monitoring*

OpenPTrack [19] is an open-source multi- $RGB$ - $D$  camera person tracking software. While this does not detect the skeleton, it is possible to monitor the trajectory of more than six people as compared to the Kinect which can only track up to six people over a large area using calibrated networked cameras. Streaming of the tracking data is done using UDP and NDN in JSON format. The RoomAlive Toolkit [11] uses multiple calibrated Kinect cameras for dynamic projection mapping to enable immersive augmented experience using projectors. RoomAlive uses a distributed network for tracking touch detection and body movements using optical-flow-based particle tracking.

Torres-Solis and Chau [24] used a combination of computer vision and dead reckoning for indoor localization. Their system is composed of wearables and fiducial markers for tracking. Klingbeil and Wark [13] demonstrated a wireless sensor network for positioning and monitoring human motion in an indoor environment. They used an inertial sensor along with a mobile sensor node worn by the person moving inside the building. Motion data is processed on the onboard mobile node and transferred to a static network of seed nodes. Based on the person's pedometric mapping, seed node position, and indoor map information, location is calculated using a Monte Carlo-based algorithm. Chen et al. [6] proposed an intelligent video monitoring system to improve the safety of old persons who have dementia. They used 23 cameras to record daily activity. Elopement activity is detected using a hidden Markov model.

Even though there are many multiple Kinect activity monitoring approaches, these systems mostly make use of Kinect v1 sensor because up to four sensors can be connected to the same PC reducing the complexity of the system. Our system, on the other hand, uses multiple Kinect v2 sensors using distributed network because of



its reasonably high accuracy, low interference between multiple cameras, and better occlusion handling. We do not use any wearable devices or fiducial markers for localization and monitoring as it has a significant impact on acceptability, especially for older people.

### 2.3 Calibration of Cameras Using RoomAlive Toolkit

RoomAlive Toolkit [11] is an open-source SDK developed by Microsoft for dynamic projection mapping research and has been in use at Microsoft Research for many years. RoomAlive Toolkit has been used by Microsoft for many interactive projection mapping and augmented reality projects. The basic building blocks of RoomAlive Toolkit consist of an ensemble of projectors and cameras, or ‘ProCam’ unit and enables developers to calibrate multiple Kinect v2 sensors and video projectors connected over the network. RoomAlive captures and creates a unified 3D model of the geometry and appearance of the room. RoomAlive Toolkit consists of two separate projects developed in C# language.

- ProCamCalibration—This project consists of KinectServer, ProjectorServer, and CalibrateEnsemble applications which are used to calibrate multiple projectors and Kinect v2 cameras.
- RoomAliveToolkitForUnity—RoomAlive Toolkit for Unity contains a set of scripts to enable dynamic projection mapping based on the calibration data from ProCamCalibration. It also streams Kinect data to Unity. This thesis research only makes use of the streaming script from this project.

RoomAlive consists of multiple projector-camera units called ProCam. Projector projects out on to the room, and Kinect camera also looks out on to that room. Projector and all the depth cameras that can see some part of the gray code that projector displays form a projector group. Multiple projector groups will be needed to create a full 3D model of a room. Another projector someplace else in the room might have two more Kinect sensors next to it, and that would form another projector group, these two groups are a distinct group and there is no way to relate the geometry of both the groups. To relate the geometry of one group to the other, a third camera which will overlap both the groups is needed. This third camera belongs to both the groups will help establish the geometry of the two groups together. Overlap of multiple camera groups is referred to as an ensemble.

Geometric camera calibration estimates the intrinsic and extrinsic of all the cameras in the ensemble. Intrinsic camera parameters provide information about the camera focal length, optical center, and image sensor format. Extrinsic camera parameters describe the coordinate system transformations from 3D world coordinates to 3D camera coordinates such as position and orientation of the camera center in world coordinates. Using intrinsic and extrinsic parameters, complete camera matrix can be derived. Complete camera matrix can be used to associate 2D points on the image

plane with 3D points in the world coordinate system. Complete camera matrix is used in applications such as background blur (limited depth of field effect) seen in recent mobile phones equipped with two cameras which uses stereo vision to calculate the 3D world coordinates of the point viewed by both cameras. Calibration is essential in computer vision systems where more than one camera are used and where exact geometrical measurements are performed. In this thesis research, the first Kinect camera establishes the coordinate system for the entire ensemble. The pose of the first Kinect camera in the ensemble will become the origin of world coordinate system, and every other camera will be calibrated to the first Kinect camera. A simple pinhole camera model can be used to describe the intrinsic, extrinsic, and complete camera matrix.

If the camera is oriented arbitrarily and does not have its center of projection at  $(0, 0, 0)$ , we need a translation and rotation to make the camera coordinates system coincide with the world coordinate system. Let the rotation be given by a  $3 \times 3$  rotation matrix  $R$  and translation by  $T(T_X, T_Y, T_Z)$ . The matrix formed by first applying translation and then rotation is given by the  $3 \times 4$  matrix as follows:

$$E = (R|RT)$$

$E$  is the ‘extrinsic matrix.’ The complete camera transformation is given by

$$K(R|RT) = (KR|KRT) = KR(I|T)$$

Therefore,  $P_C$  is given by

$$P_C = KR(I|T)P = CP$$

where  $C$  is a  $3 \times 4$  matrix usually called the ‘camera matrix.’ Because  $C$  is a  $3 \times 4$  matrix,  $P$  needs to be in 4D homogeneous coordinates and resulting  $P_C$  will be in 3D homogeneous coordinates. Exact 2D coordinates of the projection on the image plane can be obtained by dividing the first two coordinates of  $P_C$  by the third coordinate.

### 2.3.1 Calibration Process

Calibration finds the pose and position of every Kinect camera in the ensemble along with lens distortion and focal length. Calibration process consists of two phases: ‘acquisition phase’ and ‘solve phase.’ The calibration process is completely automatic and does not require user interventions. Cameras are placed in the room such that it will extend the field of view while still having ~10% overlap between the cameras.

### 2.3.2 Acquisition

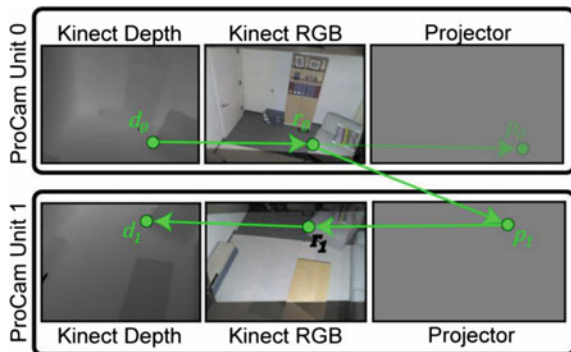
In acquisition phase, the projector displays a series of gray codes which are observed by all the color cameras and takes a snapshot by each camera. These gray codes are used to establish the mapping from a pixel in the color camera to a pixel coordinate in the projected display. Camera intrinsics such as focal length and lens distortion are obtained and stored in the XML file. It also stores the depth image and mean depth image for the calibration process.

### 2.3.3 Solving

The calibration begins with transforming each correspondence into depth image using API provided by Kinect SDK. This results in 2D to 3D point correspondences. These correspondences between 3D points and their projections on the camera images are then passed to OpenCV’s `calibrateCamera` function to find intrinsic and extrinsic parameters of each unit. To increase the robustness, RANSAC procedure is used. To find the global extrinsic between the two Kinect cameras, correspondence between two units is computed as shown in Fig. 2.1. First, the depth pixel in Unit 0 ( $d_0$ ) is mapped to an RGB pixel in Unit 0 ( $r_0$ ), and then, the corresponding projector pixel ( $p_1$ ) is mapped in Unit 1 by decoding the gray codes. Gray code correspondence is then inverted to look up the RGB pixel ( $r_1$ ) in Unit 1. Finally, the correspondence between two units is found by inverting the transfer map resulting in depth pixel ( $d_1$ ).

The solve phase also merges the color and the depth data collected from all the Kinect sensors and creates a unified 3D model of the room as shown in Fig. 2.2, which helps the user to check the quality of registration. The model is also exported to an obj file which can be directly imported into Unity.

**Fig. 2.1** Establishing global extrinsic by chaining together correspondences





**Fig. 2.2** A unified 3D model of the room generated during calibration

### 2.3.4 Code Components

The RoomAlive Toolkit code is completely written in C#. This section discusses four components used for the calibration process: (1) *ProCamEnsembleCalibration.dll*; (2) *KinectServer*; (3) *ProjectorServer*; and (4) *CalibrateEnsemble*.

#### 2.3.4.1 *ProCamEnsembleCalibration.dll*

*ProCamEnsembleCalibration.dll* is the core of the project and has all the calibration codes in it. It is used in the calibration step to acquire calibration patterns and calculate the camera parameters of Kinect and projector such as focal length, optical center, and the dimensions of the camera. It calculates the pose, i.e., position and orientation of all the Kinects and the projectors. It provides ways to interact and pull out the calibration results at runtime. *ProjectorCameraEnsemble* is the main class that uses other helper classes to check for the connected cameras and projector, makes the projector display the gray code, and captures them with Kinect's color camera. Then, it decodes the gray code and finds the pose of Kinects and projectors. It also creates the unified 3D model of the room and saves it to .obj file. The *KinectServer2Client* and *ProjectorServerClient* classes are used to communicate with their relative servers. The *Kinect2Calibration* class does all the calibration of the Kinect sensor. *GrayCode* class is used to create gray code images using *ARGBImage*. The *Matrix* class handles the matrix calculation.

### **2.3.4.2 KinectServer**

Single Kinect v2 sensor can be connected to a computer simultaneously; therefore, KinectServer runs on every PC that has a Kinect sensor. The KinectServer is a tool to distribute color data, depth data, and camera intrinsic from each Kinect cameras around the network. It is a stand-alone application with no user interface. It uses Windows Communication Foundation (WCF) and requires Kinect for Windows V2 SDK.

### **2.3.4.3 ProjectorServer**

The ProjectorServer is analogous to KinectServer. It runs on every PC that has a projector connected to it. During the acquisition phase, each projector is assigned to an index specified in the XML file and displays the gray code from the projector. Projector server is only required during the acquisition phase of the calibration setup.

### **2.3.4.4 CalibrateEnsemble**

The CalibrateEnsemble provides the GUI for the calibration process. It also assists in setup and configuring all the Kinect sensors and projectors for calibration. It creates an XML file to store the calibration information and allows the user to view the results.

## **2.4 Hardware Setup**

This section discusses the overall hardware setup and necessary initialization of all the Kinect sensors and projectors used in the system.

### ***2.4.1 Hardware Specification***

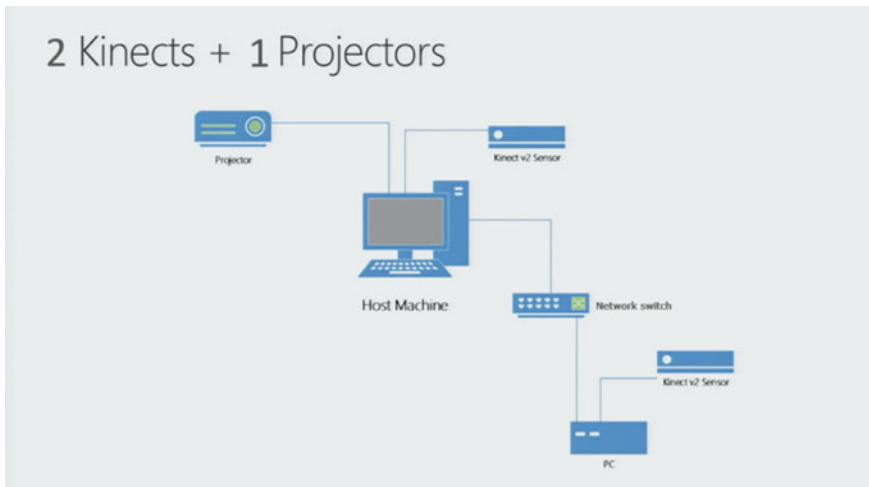
In this project, only one ProCam unit is used. ProCam unit was created using two Kinect sensors and one projector. Optoma TX536 projector was used to calibrate the two Kinects. Optoma TX536 projector does not have a wide field of view and only covers a small portion of the room, but it is sufficient for calculating the position and orientation of the Kinect. Wider field of view of the projector will result in more accurate intrinsic calculations of the Kinect cameras. The two Kinect v2 cameras were placed horizontally with a slight tilt of around  $5^\circ$ . The ProCam unit's connection diagram is shown in Fig. 2.3.

The host computer was an Intel Core i7-6820HQ 2.7 GHZ CPU with 8 GB DDR3 RAM running Windows Server 2016. First, Kinect v2 sensor is connected to the host PC via USB 3.0. Projection is also connected to the host PC because ProjectorServer.exe does not support distributed rendering framework. The second Kinect sensor is connected to another PC with Intel Core i7-6600U 2.5 GHZ CPU and 8 GB DDR3 RAM running Windows 10. At first, both the PCs were connected using Wi-Fi to the network switch which significantly slows down the data transfer rate between the PCs. Later, we switched to a more stable and faster wired ethernet connection.

### 2.4.2 ProCam Unit Placement

The primary goal of Kinect placement is to increase the field of view. A different arrangement of Kinect sensors in the room has been tried for successful calibration. Requirements for proper room setup are as follows: (1) The Kinect sensors should be placed such that both the sensors should view the gray codes projected by the projector. At least 10% overlap should be there; (2) precise alignment of the Kinect sensors is not required; (3) both the color camera and the depth camera of both the Kinect sensors must observe the projected gray code; (4) the projector should be configured in 'desktop front' projection mode; (5) Windows should be set to 'Extend' its desktop to the projector.

Figure 2.4 displays the position of the two Kinect sensors and projector in the room.



**Fig. 2.3** Hardware connection of ProCam unit



Fig. 2.4 Kinect sensors and projector placement in the room

### 2.4.3 *Configuring Calibration.xml*

CalibrateEnsemble.exe has a user interface which allows the user to select the number of cameras and projectors in the ensemble and automatically creates a new calibration XML file with some of the information filled in. XML file has a number of cameras and projectors. The user can rename the cameras and projectors listed in the XML file. We have to manually enter the hostname or IP address for each Kinect cameras and projectors in the XML file. The pose information is a  $4 \times 4$  matrix which is by default set to identity matrix for the first Kinect in the XML file. That means the first camera is in that pose within the global coordinate frame. This default pose matrix of the first camera will not change post calibration. Alternatively, the user can establish a different position for that camera manually. We also have to change the 'displayindex' value under the projector. For an external display such as a projector, the displayindex value should be set to '1', '0' being the main display.

### 2.4.4 *Practical Issues*

When placing the objects in the scene for the projector to project on, we had to make sure that there were no black objects in the scene. Kinect v2 sensor uses infrared time-of-flight technology to detect depth in the scene. Black objects absorb most of the infrared light, and hence, most of the infrared laser light transmitted by the Kinect sensor was getting absorbed by the black objects resulting in corrupted depth data and unsuccessful calibration.

The best Kinect sensor placement to avoid occlusion is to place the sensor higher than the average human height or mount the sensors on the ceiling. In both of the cases, the sensor needs to be tilted downward. If the first Kinect sensor is placed tilted, the global coordinate system will also have a tilt as it is aligned with the first sensor. The global coordinate system can be corrected, so the gravity points downward by using Kinect's accelerometer data. In our hardware setup, the tilt of the first sensor is kept close to  $0^\circ$ .

## 2.5 Software Implementation

In this section, we describe the implementation of human trajectory and activity tracking based on the Kinect skeleton data received by the host PC from two calibrated Kinect cameras. The system stores the skeleton data in a CSV file for further processing and categorizes the data based on the different area of the room and whether the person is sitting or standing. It calculates the total time spent by the person in each category. Finally, it displays the total time spent data in the form of a pie chart.

### 2.5.1 *Environment Setup*

The language used for programming is C# throughout the implementation. The application is developed in Unity 2017 game development engine using Microsoft Visual Studio Community version as an editor. Kinect for Windows SDK V2.0 color and depth image streams were used for proper placement of Kinect sensors. Revision control is done using Git. Revision control is used to manage all the iterative changes of program. GitHub is used for revision control of the software code. It allows the user to revert to any previous version of the code easily and to keep track of all the changes committed to the code.

### 2.5.2 *User Interface*

The user interface includes a 3D model of the room with the two Kinect sensor 3D models placed at the exact locations in the room as determined by the calibration. 'Start Recording' button records full skeleton data into a CSV file. The user can stop the recording of skeleton data by clicking on 'Stop Recording' button. 'Process data' button in the UI once clicked reads back the stored skeleton data, classifies the data and calculates the total time spent in each category, and draws a pie graph using 'Graph and Chart' plugin. It also displays the total time spent in each category. The user interface of the application is shown in Fig. 2.5.

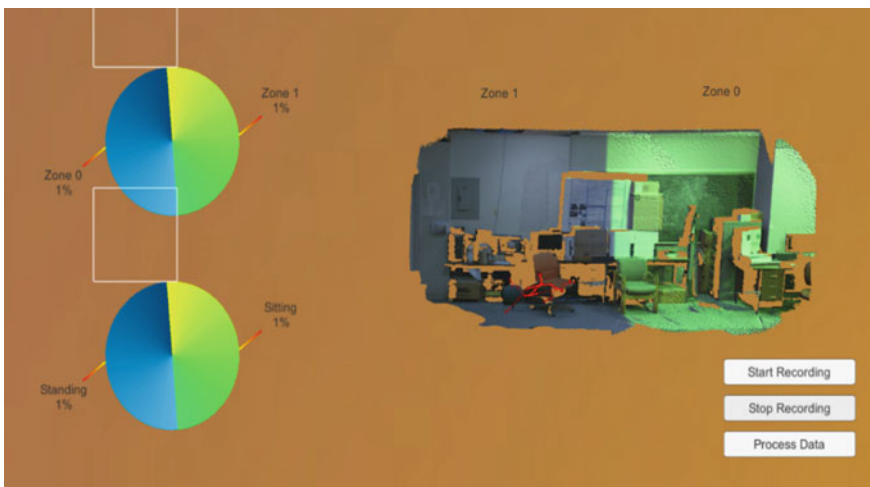


### 2.5.3 Creating 3D Model

The calibration process generates .xml and .obj files which is used in Unity to create the 3D model of the room. Object file describes the geometric properties of the room and can be directed imported in Unity Engine. Photographs captured by both the Kinect cameras are used to colorize the 3D model. Each Kinect sensor is linked to their respective empty game object. XML calibration file contains the pose matrix of the Kinect sensors which are used to position the Kinect sensors game object in the 3D model. 'RATCalibrationData' script reads the XML file to extract pose information. 'RATSceneSetup' contains the helper functions that automate the connection and dependencies among many game objects. All the calibration files, 'RATCalibrationData' and 'RATSceneSetup', are linked to an empty game object, and *Build RoomAlive Scene* button in 'RATSceneSetup' script will create a complete scene in that object.

### 2.5.4 World Coordinate System

As described earlier, the coordinate system of the 3D model is centered at the first Kinect sensor in the ensemble. The first sensor is positioned at (0, 0, 0), and every other object in the scene is translated from the first Kinect sensor. The calibration data provided by the RoomAlive Toolkit uses a right-handed coordinate system similar to Kinect sensor, while Unity uses a left-handed coordinate system. Real-world positions and the RoomAlive Toolkit's coordinate system are one to one.

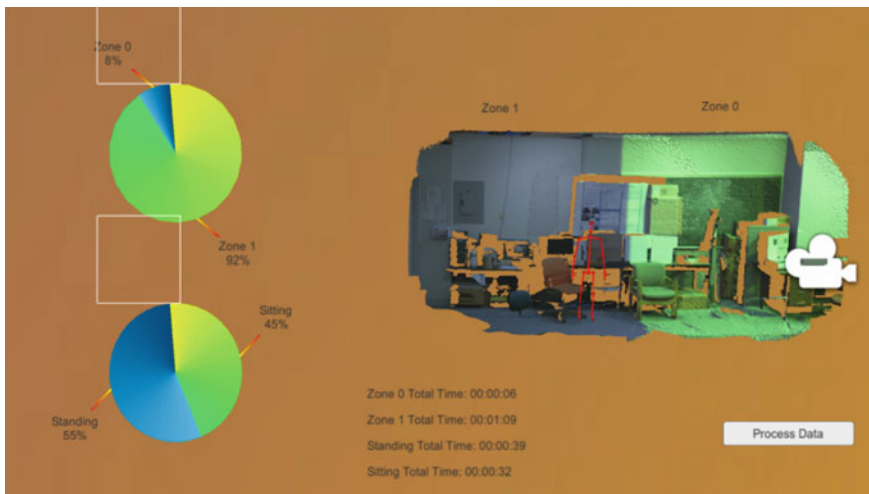


**Fig. 2.5** User interface of Kinect Tracker

### 2.5.5 User Tracking

Skeleton data from all the Kinect sensors is transferred to the host computer using Windows Communication Foundation where the main application runs. All the skeletons along with their tracking states are stored in RATKinectSkeleton data container. Each Kinect sensor can track up to six skeletons, for two Kinect sensors that are total of twelve tracked skeleton. RATKinectSkeleton merges the skeletons data from all the sensors making total tracked skeletons up to six. The skeleton data is in the coordinate system of the tracking Kinect sensor. All the skeleton data received by the host computer is assumed to be tracked by the main Kinect sensor. If the skeleton data is detected by the second Kinect sensor, to display the skeleton data in the 3D model, the 3D point needs to be converted into world coordinate system using the localToWorldMatrix, where the localToWorldMatrix is derived from the pose of the second camera. Each tracked 3D point is multiplied with transform.localToWorldMatrix where the transform is attached to the game object of the respected Kinect camera. Kinect sensor v2 detects 25 joints, and all the joint data is transferred to the host computer. In this thesis research even though all the tracked joints are used to display on the user interface, only 21 joints of the tracked person are saved in the CSV file and to do all the data processing. We can also extract the height information of the tracked person to determine whether the person is sitting or standing. Figure 2.6 shows the skeleton data overlaid on the 3D model using Gizmos.

Gizmos are used to overlay the skeleton data on the 3D model. Gizmos in Unity are used for visual debugging and can only be seen in scene view or by enabling Gizmos in game view. Gizmos contain many static methods to draw different shapes such as a cube, line, mesh, and sphere.



**Fig. 2.6** Skeleton data overlaid on the 3D model

### **2.5.6 Recording Data**

The application starts recording the skeleton data once ‘Start Recording’ button is clicked. All the joints for every detected skeleton are checked for the tracking state of the joint. If the joint is tracking state is tracked, the world coordinates of the joint along with the current time stamp are passed to ‘Save’ method of CSV class which saves the data in CSV file. The time stamp unit is in millisecond since midnight.

### **2.5.7 Activity Classification Model**

The raw skeleton data along with time stamp is stored in a CSV file. This data is read back and classified into two different categories as follows: (1) zones and (2) standing/sitting. The first category classifies the data into two different zones defined in the 3D model. 3D model is created using two cameras, and each zone indicates area viewed by their respected camera. Zone 0 is the area monitored by the first Kinect camera, and Zone 1 is the area monitored by second Kinect camera in the ensemble. The total area being monitored is segmented into two zones by setting a threshold of  $X$ -coordinate precisely at the center of both the cameras. The first camera’s  $X$ -coordinate is 0, and the  $X$ -coordinate of the second camera represents the horizontal distance between both the cameras. The threshold is set by dividing the  $X$ -coordinate of the second camera by two.

The application also classifies the data depending on whether the tracked human is standing or sitting based on the height of the head joint. The threshold which decides whether the person is sitting or standing is hardcoded in the code based on the room geometry and Kinect sensor placement. The raw data is read back from the CSV file after clicking ‘process data’ button on the UI and stored into ‘CompleteDataList’ List for classification. The code then iterates through all the elements in ‘CompleteDataList’ list, segments the data according to two categories as described earlier, and stores the data into respected CSV file and list of each activity.

The total time spent in each category is calculated and displayed on the user interface. The total time is calculated by subtracting the current frame time from the next frame time and adding it to ‘TotalTimeStanding’ variable. If the time difference between two consecutive frames is greater than 2 s, it is discarded.

### **2.5.8 Graphs**

Graphs are used to compare the total time spent in two categories efficiently. Two pie charts are used to compare two categories: one for Zone 0 and Zone 1, and another pie chart for standing and sitting. ‘Graphs and Charts’ plugin is used to draw pie

charts. Once ‘Graphs and Charts’ plugin is imported as an asset in the project, a pie chart can be easily added to the scene as a game object. Before setting values, the total time is converted into a percentage and rounded to the nearest value. The categories for the pie chart are defined in the inspector window.

## 2.6 Experimental Result

This section describes the qualitative results that demonstrate the performance of the system in a highly complex environment with occlusion. In addition, some of the practical scenarios where the dataset created by the system can be used are discussed.

Figure 2.7 presents the scatter plot of the projected path traveled in world coordinates over a period in the test room as recorded by our application. The origin of the world coordinate system is located at the first Kinect camera in the ensemble. The

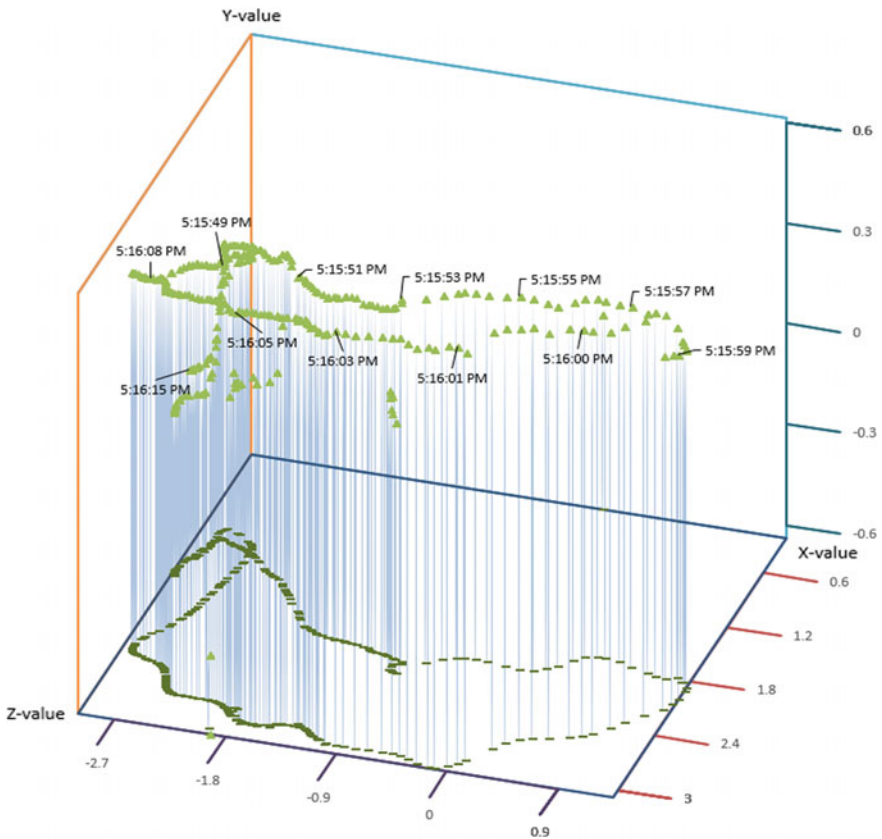


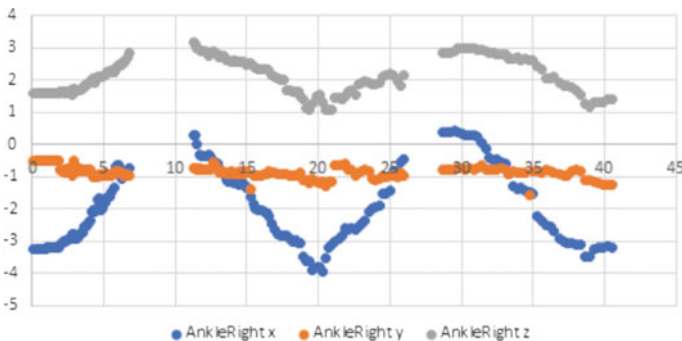
Fig. 2.7 Trajectory of the tracked person in the world coordinate system

first Kinect sensor is positioned at 0 on the  $X$ -axis in the world coordinate system, and the second Kinect sensor is positioned at  $-2.26$  on the  $X$ -axis in the world coordinate system. The world coordinate system in the tracking application is the same as the coordinate system of Kinect SDK. The scatter plot represents the motion capture over the period of approximately 25 s. The second Kinect sensor is directly connected to the host computer where our application runs, and the first Kinect sensor is connected to another computer and sends the skeleton data over the network to the host computer. Due to this networked configuration, the frame rate of the skeleton data of first Kinect sensor is significantly lower than the frame rate of the skeleton data from the second Kinect sensor. The difference in the frame rate between both the sensors can be noticed in the scatter plot as the area covered by the first Kinect sensor consists of fewer number of markers as compared to the area covered by the second Kinect sensor.

The current implementation consisting of tracking, logging, and processing the skeleton data operates at around 12 frames/second. Since the small motion of 60 s captures database of 720 examples, a linear search for the corresponding joint position is conducted. For the considerably larger database, clustering would allow for faster processing.

Figure 2.8 shows the dataset of the person walking with respect to time across the room twice as represented by two peaks in the  $X$ -coordinate position of the joint. The gaps in the dataset denote the time required by the second Kinect sensor to detect the skeleton. By the time the person walking across the room was in the field of view of the second Kinect, he was almost facing his back toward the second camera and got self-occluded. This occlusion problem can be eliminated by using multiple cameras to cover the same area.

The skeleton data recorded by this system can also be used for other applications such as gesture recognition, posture analysis, and gait analysis. Gait analysis can be done using angular kinematics on the skeleton data recorded by the application. This system provides continuous skeleton tracking data over a large area which is required to do a gait analysis on the dataset. Furthermore, machine learning can be applied to gait analysis data for disease identification.



**Fig. 2.8** Walking dataset as recorded by the application

## 2.7 Conclusion

Assisted living technology is shifting from care facilities to patient's home due to low-cost tracking solutions. Multiple inexpensive depth cameras such as Kinect connected over the network can cover every part of the house and provide a person's location and behavioral data, enabling better care for seniors who have chronic diseases. We have described a system that can track and record the skeleton data of people in the smart indoor environment. The system consists of multiple calibrated Kinect depth sensors for tracking and localization. The system was able to achieve real-time positioning, path recording, and monitoring behavioral information using markerless computer vision tracking. It can detect humans in the dark, invariant to background lighting condition, clothing, and skin color. Finally, the behavioral data is stored in a CSV file and the total time spent in each category is calculated and displayed on the user interface in text and graphical form.

The classification model is designed to classify the activity of only one person. Future work involves creating a separate data container for every tracked person and storing the classification data of every tracked skeleton separately. Recording frame rate can be improved further by storing multiple frames data at once in CSV file. The behavior of the Kinect skeleton tracker is not always perfect, and hence, the abrupt changes in skeleton data need to be filtered. Future work also includes tracking and reidentification of multiple users using weighted bone length and particle filtering. Furthermore, this system is designed to work in a single large area with numerous Kinects and a projector with an overlapping field of view creating an ensemble. For multiple rooms, multiple ensembles should be incorporated in a single application running on the host computer.

## References

1. Atrsaei A, Salarieh H, Alasty A, Abediny M (2018) Human arm motion tracking by inertial/magnetic sensors using unscented Kalman filter and relative motion constraint. *J Intell Rob Syst* 90(1–2):161–170
2. Azis NA, Choi HJ, Iraqi Y (2015) Substitutive skeleton fusion for human action recognition. In: 2015 International conference on big data and smart computing (BigComp). IEEE, pp 170–177
3. Beymer D, Konolige K (1999) Real-time tracking of multiple people using continuous detection. In: IEEE Frame rate workshop, pp 1–8
4. Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Van Gool L (2009) Robust tracking-by-detection using a detector confidence particle filter. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 1515–1522
5. Caon M, Yue Y, Tscherrig J, Mugellini E, Khaled OA (2011) Context-aware 3d gesture interaction based on multiple kinects. In: Proceedings of the first international conference on ambient computing, applications, services and technologies, AMBIENT. Citeseer, pp 7–12
6. Chen D, Bharucha AJ, Wactlar HD (2007) Intelligent video monitoring to improve safety of older persons. In: 2007 29th Annual international conference of the IEEE engineering in medicine and biology society. IEEE, pp 3814–3817
7. Comaniciu D, Ramesh V, Meer P (2003) Kernel-based object tracking. *IEEE Trans Pattern Anal Mach Intell* 5:564–575

8. Ercan AO, Gamal AE, Guibas LJ (2013) Object tracking in the presence of occlusions using multiple cameras: a sensor network approach. *ACM Trans Sensor Netw (TOSN)* 9(2):16
9. Fuentes LM, Velastin SA (2001) People tracking in surveillance applications. In: *Proceedings of 2nd IEEE international workshop on PETS, Kauai, Hawaii, USA*
10. Ikemura S, Fujiyoshi H (2010) Real-time human detection using relational depth similarity features. In: *Asian conference on computer vision*. Springer, Berlin, pp 25–38
11. Jones B, Sodhi R, Murdock M, Mehra R, Benko H, Wilson A, Ofek E, MacIntyre B, Raghuvanshi N, Shapira L (2014) Roomalive: magical experiences enabled by scalable, adaptive projector-camera units. In: *Proceedings of the 27th annual ACM symposium on user interface software and technology*. ACM, pp 637–644
12. Kalman RE (1960) A new approach to linear filtering and prediction problems. *J Basic Eng* 82(1):35–45
13. Klingbeil L, Wark T (2008) A wireless sensor network for real-time indoor localisation and motion monitoring. In: *2008 International conference on information processing in sensor networks (IPSN 2008)*. IEEE, pp 39–50
14. Lin Q, Zhang D, Chen L, Ni HB, Zhou S (2014) Managing elders' wandering behavior using sensors-based solutions: a survey. *Int J Gerontol* 8(2):49–56
15. Lun R, Zhao W (2015) A survey of applications and human motion recognition with microsoft kinect. *Int J Pattern Recognit Artif Intell* 29(05):1555008
16. Lun R, Gordon C, Zhao W (2016) The design and implementation of a kinect-based framework for selective human activity tracking. In: *2016 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, pp 002890–002895
17. Lun R, Gordon C, Zhao W (2016) Tracking the activities of daily lives: an integrated approach. In: *2016 Future technologies conference (FTC)*. IEEE, pp 466–475
18. Masuyama G, Kawashita T, Umeda K (2017) Complementary human detection and multiple feature based tracking using a stereo camera. *ROBOMECH J* 4(1):24
19. Munaro M, Basso F, Menegatti E (2016) Openprack: open source multi-camera calibration and people tracking for RGB-D camera networks. *Robot Auton Syst* 75:525–538
20. Papoulis A, Pillai SU (2002) *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education
21. Poland MP, Nugent CD, Wang H, Chen L (2012) Genetic algorithm and pure random search for exosensor distribution optimisation. *Int J Bio-Inspired Comput* 4(6):359–372
22. Ponraj G, Ren H (2018) Sensor fusion of leap motion controller and flex sensors using Kalman filter for human finger tracking. *IEEE Sens J* 18(5):2042–2049
23. Spinello L, Arras KO (2011) People detection in RGB-D data. In: *2011 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, pp 3838–3843
24. Torres-Solis J, Chau T (2010) Wearable indoor pedestrian dead reckoning system. *Pervasive Mob Comput* 6(3):351–361
25. Williamson B, LaViola J, Roberts T, Garrity P (2012) Multi-kinect tracking for dismounted soldier training. In: *Proceedings of the interservice/industry training, simulation, and education conference (I/ITSEC)*, pp 1727–1735
26. Wren CR, Pentland AP (1998) Dynamic models of human motion. In: *Proceedings third IEEE international conference on automatic face and gesture recognition*. IEEE, pp 22–27
27. Zhao W, Lun R, Espy DD, Reinthal MA (2014) Realtime motion assessment for rehabilitation exercises: integration of kinematic modeling with fuzzy inference. *J Artif Intell Soft Comput Res* 4(4):267–285
28. Zhao W, Lun R, Gordon C, Fofana ABM, Espy DD, Reinthal A, Ekelman B, Goodman GD, Niederriter JE, Luo C et al (2016) Liftingdoneright: a privacy-aware human motion tracking system for healthcare professionals. *Int J Handheld Comput Res (IJHCR)* 7(3):1–15
29. Zhao W, Lun R, Gordon C, Fofana ABM, Espy DD, Reinthal MA, Ekelman B, Goodman GD, Niederriter JE, Luo X (2017) A human-centered activity tracking system: toward a healthier workplace. *IEEE Trans Human-Mach Syst* 47(3):343–355
30. Zhao W, Reinthal MA, Espy DD, Luo X (2017) Rule-based human motion tracking for rehabilitation exercises: realtime assessment, feedback, and guidance. *IEEE Access* 5:21382–21394

31. Zhao W, Wu Q, Reinthal A, Zhang N (2018) Design, implementation, and field testing of a privacy-aware compliance tracking system for bedside care in nursing homes. *Appl Syst Innov* 1(1):3
32. Zhao W (2016) A concise tutorial on human motion tracking and recognition with Microsoft kinect. *Sci China Inf Sci* 59(9):93101
33. Zhao W (2016) On automatic assessment of rehabilitation exercises with realtime feedback. In: 2016 IEEE international conference on electro information technology (EIT). IEEE, pp 0376–0381
34. Zhao W, Lun R (2016) A kinect-based system for promoting healthier living at home. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 000258–000263
35. Zhao W, Espy DD, Reinthal MA, Feng H (2014) A feasibility study of using a single kinect sensor for rehabilitation exercises monitoring: a rule based approach. In: 2014 IEEE symposium on computational intelligence in healthcare and e-health (CICARE). IEEE, pp 1–8
36. Zhao W, Feng H, Lun R, Espy DD, Reinthal MA (2014) A kinect-based rehabilitation exercise monitoring and guidance system. In: 2014 IEEE 5th international conference on software engineering and service science. IEEE, pp 762–765
37. Zhao W, Lun R, Espy DD, Reinthal MA (2014) Rule based realtime motion assessment for rehabilitation exercises. In: 2014 IEEE symposium on computational intelligence in healthcare and e-health (CICARE). IEEE, pp 133–140
38. Zhao W, Espy DD, Reinthal MA, Ekelman B, Goodman G, Niederriter J (2015) Privacy-aware human motion tracking with realtime haptic feedback. In: 2015 IEEE international conference on mobile services. IEEE, pp 446–453
39. Zhao W, Lun R, Gordon C, Fofana AB, Espy DD, Reinthal MA, Ekelman B, Goodman G, Niederriter J, Luo C et al (2016) A privacy-aware kinect-based system for healthcare professionals. In: 2016 IEEE international conference on electro information technology (EIT). IEEE, pp 0205–0210
40. Zhao W, Wu Q, Espy DD, Reinthal MA, Luo X, Peng Y (2017) A feasibility study on using a kinect-based human motion tracking system to promote safe patient handling. In: 2017 IEEE international conference on electro information technology (EIT). IEEE, pp 462–466
41. Zhao W, Wu Q, Padaraju V, Bbela M, Reinthal A, Espy D, Luo X, Qiu T (2017) A privacy-aware compliance tracking system for skilled nursing facilities. In: 2017 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 3568–3573
42. Zhao W, Pillai JA, Leverenz JB, Luo X (2018) Technology-facilitated detection of mild cognitive impairment: a review. In: 2018 IEEE international conference on electro/information technology (EIT). IEEE, pp 0284–0289
43. Zhu L, Wong KH (2013) Human tracking and counting using the kinect range sensor based on Adaboost and Kalman filter. In: International symposium on visual computing. Springer, Berlin, pp 582–591



# Chapter 3

## Unobtrusive Sensing Solution for Post-stroke Rehabilitation



**Idongesit Ekerete, Chris Nugent, Oonagh M. Giggins  
and James McLaughlin**

**Abstract** This Chapter proposes an unobtrusive sensing solution for monitoring post-stroke rehabilitation exercises within a home environment. It begins with the definition of stroke, its types, statistics and effects. An overview of stroke rehabilitation techniques ranging from multiple exercising and isolated approaches to motor skill learning, mirror imagery, adjuvant therapies and technology-based interventions are all presented in this Chapter. In addition, the potential for the use of unobtrusive sensing solutions such as thermal, radar, optical and ultrasound sensing are considered with practical examples. The Seebeck, time of flight (ToF) and Doppler principles, which are associated with a number of the sensing solutions, are also explained. Furthermore, sensor data fusion (SDF) and its architectures such as centralized, distributed and hybrid architectures are explained. A few examples of SDF applications in automobile and terrestrial light detection are included in addition to the advantages and disadvantages of the approaches. Unobtrusive sensing solutions and their applications in healthcare are captured in this Chapter. The Chapter includes details of initial experimental results on post-stroke rehabilitation exercises which were obtained using thermal and radar sensing solutions. The Chapter concludes with an outline of recommendations for future research.

**Keywords** Rehabilitation · Post-stroke · Unobtrusive · Wearable · Radar · Thermal · Sensors

---

I. Ekerete (✉) · C. Nugent  
School of Computing, Ulster University, Newtownabbey BT37 0QB, Northern Ireland, UK  
e-mail: [ekerete-i@ulster.ac.uk](mailto:ekerete-i@ulster.ac.uk)

C. Nugent  
e-mail: [cd.nugent@ulster.ac.uk](mailto:cd.nugent@ulster.ac.uk)

O. M. Giggins  
Dundalk Institute of Technology, NetwellCASALA, Dundalk, Republic of Ireland  
e-mail: [Oonagh.Giggins@dkit.ie](mailto:Oonagh.Giggins@dkit.ie)

J. McLaughlin  
NIBEC, Ulster University, Newtownabbey BT37 0QB, Northern Ireland, UK  
e-mail: [jad.mclaughlin@ulster.ac.uk](mailto:jad.mclaughlin@ulster.ac.uk)

## 3.1 Introduction

Nowadays, many technology-based products and services are embedded with sensors. Some of which are designed to be able to count, measure, guide or manipulate chemical, physical and biological processes [1]. However, data acquisition formats differ from sensor to sensor. While some interact directly with a system to acquire data, effect changes and manage specified conditions, others perform these functions remotely without any physical contact with the system. In the medical and health sectors, for instance, sensing solutions such as powered wheelchair, surgical mask, heart valves, neuromuscular stimulator, ultrasound and magnetic resonance imaging machines, amongst others carry out data acquisition by either of these methods [2, 3].

Data acquisition for process monitoring during rehabilitation can give valuable information to help make an informed decision on prescription and treatment; implement corrective measures and analyses, and evaluate the workability of an applied solution [4]. Furthermore, physiotherapists and occupational therapists benefit from sensing solutions to monitor recovery post-stroke [5]. One way is through data acquired and displayed by video cameras and force plates. These help to estimate correct posture and gait during upper and lower extremities rehabilitation exercises [5]. Other parameters such as speed, direction of motion and distance covered can also be estimated using an appropriate sensing solution [6, 7].

In addition, scientific monitoring processes can be carried out using wearable sensors and video cameras at rehabilitation centres and hospitals [4]. While the usefulness of these centres and the advantages of wearable monitoring devices cannot be over-emphasized, factors such as transportation constraints, limited personal budget and unavailability of appointment slots are some of the problems of visiting the centres and hospitals, especially, if they are situated at a distant location from the user. Moreover, wearable devices and video cameras pose problems ranging from wearability and battery life to a feeling of discomfort and privacy concerns [8]. This chapter, therefore, proposes the use of a non-charged, non-wearable, non-obtrusive and a privacy-enhanced sensing solution aimed at monitoring post-stroke rehabilitation exercises within a home environment using a combination of thermal and radar sensing solutions.

## 3.2 Concept of Stroke

### 3.2.1 Definition of Stroke

The world health organization (WHO) defines stroke as the sudden onset of neurological symptoms caused by a circulatory disorder in the brain lasting for more than 24 h [9]. A stroke is further described as a brain attack which occurs when the flow of oxygenated blood to a part of the brain is cut off due to a rupture or blockage in

the brain's blood vessel. This results in damage to the brain cells in the affected area. The effect of a stroke, however, depends on the severity of the attack and the area of the brain in which the attack occurs. Some common symptoms of stroke include numbness, dysfunction of the neuromuscular functions and even death [10].

### **3.2.2 Types of Stroke**

There are two main types of strokes namely: ischemic and haemorrhagic strokes. While the former is further divided into full ischemic attack (FIA) and transient ischemic attack (TIA), haemorrhagic stroke is subdivided into intracerebral and subarachnoid haemorrhage [10].

#### **3.2.2.1 Ischemic Stroke**

Ischemic stroke occurs due to partial or total blockage in the brain's large arteries. It may result in death of brain tissues normally described as cerebral infarction. It is similar to a heart attack except that, it is only the brain's blood vessels that are involved.

##### **Full Ischemic Attack (FIA)**

Ischemic stroke incidents are mostly caused by the accumulation of obstructive substances in the cerebral blood transportation routes. These substances can include blood clots, cholesterol crystals, fat deposits and foreign bodies which emanate from degenerated biomaterials in the human body. A case of FIA is otherwise known as full stroke.

Furthermore, an area of the brain that has experienced blood dearth due to obstructions is called area of ischaemia. The aftermath effects include damage to the brain cells within the area and subsequent death of brain tissues in the affected area. This is referred to as necrosis. Ischemic stroke occurs in 80% of stroke incidents. It can be very harmful depending on the part of the brain that is affected and the time taken before intervention [11].

##### **Transient Ischemic Attack (TIA)**

A temporary obstruction to the blood supply in the brain lasting for a period less than 24 h is referred to as a transient ischemic attack (TIA). It can later result in a full-blown stroke if the mitigating factors are not properly treated. Hence, TIA is regarded as a risk factor that requires urgent intervention. TIA is normally referred

to as a ‘mini-stroke’ and should be treated as a serious incident to avoid a full-blown attack, which normally occurs in 8% of TIAs [10].

### 3.2.2.2 Haemorrhagic Stroke

Haemorrhagic stroke is derived from the word haemorrhage meaning leakage of blood from a ruptured or torn blood vessel. The broken blood vessel results in an artificial compression of brain tissues. This further heightens the spread of blood outside the vessel (haematoma). Haemorrhagic stroke is, however, of higher risk and severity than ischaemic stroke and its types are determined by the nature of bleeding: whether it is within or on the brain’s surface. The different types are explained in Sections “[Intracerebral Haemorrhagic Stroke](#)” and “[Subarachnoid Haemorrhagic Stroke](#)”.

#### Intracerebral Haemorrhagic Stroke

Deep within the brain, an arterial rupture may exist resulting in bleeding and damage to the brain tissues. This occurrence is described as intracerebral haemorrhagic stroke. Its root cause can be from age-related diseases such as arterial hardening, arteriovenous malfunction or chronic high blood pressure [12].

#### Subarachnoid Haemorrhagic Stroke

Subarachnoid haemorrhagic stroke occurs when cerebral bleeding takes place on the brain’s surface. It includes an arterial rupture on or prior to the brain causing blood retention in the subarachnoid space [11]. Some of the causes of subarachnoid haemorrhage include smoking, high blood pressure, use of oestrogenic contraceptives, drug abuse or excessive consumption of alcohol. Sudden occlusion and severe headaches are some of the common symptoms of this type of haemorrhagic stroke [12].

## 3.2.3 Stroke Statistics

### 3.2.3.1 Stroke Statistics Worldwide

Stroke is the major cause of adult disability, the third cause of death in developed countries and the second deadly disease in the world [10]. Statistics [10] worldwide have indicated that a stroke incident occurs every 2 s. 14 million first-time stroke cases were recorded in 2016 [10]. Figure 3.1 presents the 10 global causes of death in 2015 with stroke ranked second. It also indicates that more than 6 million deaths

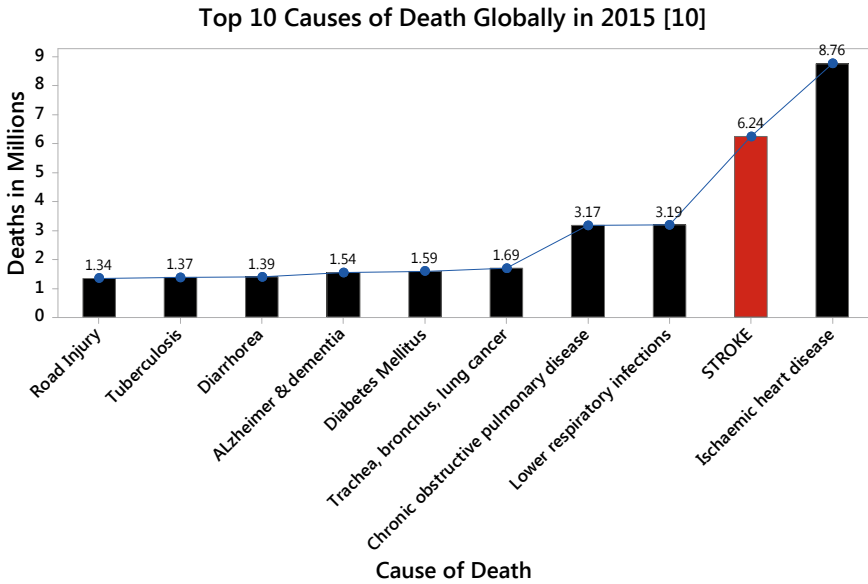


Fig. 3.1 Top ten global causes of death

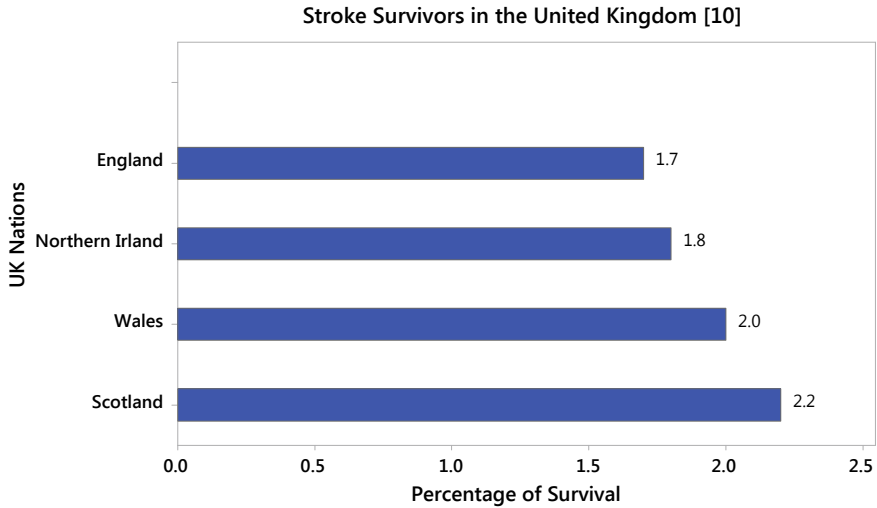
are caused annually by stroke. This figure is expected to double by the year 2035 if the risk factors are not abridged [10].

### 3.2.3.2 Stroke Incidents in the UK

Stroke incidents in the UK have been estimated at approximately 100,000 incidents per year, with more than 1.2 million survivors [10]. Figure 3.2 represents the percentage of stroke survivors in Scotland, Wales, Northern Ireland and England [10].

### 3.2.4 Symptoms of Stroke

The psychological and physical impacts post-stroke can be devastating. In the UK, for instance, a stroke campaign organization suggested that 45% of post-stroke patients felt abandoned after hospital discharge [13]. Defects in motor functions are some of the physical symptoms which are experienced post-stroke. These can include weakness of the paretic limb, oedema, subluxation, contracture, coordination and speech problems, and changes in muscle tone. Upper extremity motor dysfunctions account for 70% of post-stroke motor impairments [14].



**Fig. 3.2** Percentage of stroke survivors per country in the UK in 2015

### 3.3 Review of Stroke Rehabilitation Techniques

Clinical and research evidence suggests that the human brain can experience a significant amount of recovery after stroke with the help of an appropriate rehabilitation mechanism [15–17]. This evidence has motivated the use of a range of post-stroke rehabilitation techniques over the years such as exercise therapies, isolated concepts, motor skill learning, mirror neurons, motor imagery and adjuvant therapies [18]. A succinct review of these techniques is presented in the following Sections.

#### 3.3.1 Exercise Therapies

Exercise-based rehabilitation is a vital aspect of post-stroke rehabilitation [19]. It can be tailored to the specific needs of an individual and to achieve set goals. The key elements of exercises post-stroke are timing, intensity, duration and feedback. The latter can include visual or audio feedback [18].

##### 3.3.1.1 Bobath Approach

The Bobath concept is intended to help post-stroke sufferers regain muscle tone and motor function by focusing on specific regions of the body. Although there are indications that the Bobath concept may be useful in spasticity handling [20], however, it is not considered an effective technique for home-based rehabilitation

since it requires an active participation and physical presence of both physiotherapists and patients [18].

### 3.3.1.2 Perfetti Approach

The Perfetti approach is another exercise approach aimed at retraining the cognitive function to understand joint positions. Hatem et al. [18] reported the absence of a significant difference in motor function in patients using this method compared with those on a standard occupational therapy; hence, it is not recommended as home-based therapy technique.

### 3.3.2 Isolated Approaches

Isolated approaches are sometimes considered as integral components of exercise-based rehabilitation because, instead of considering many exercises, a specific exercise is chosen at a time. Moreover, they include interventions such as stretching, isokinetic muscle strengthening, bilateral training and forced use [18].

A systematic review by Pollock et al. [21] involving 27 studies revealed that physical rehabilitation activities like arm stretching, forced use, sit-to-stand, muscle strengthening and gait retraining offer significant improvement compared with no treatment in the recovery of motor dysfunctions. Furthermore, their gains were retained after the treatments were discontinued. On arm stretching, Hatem et al. [18] maintained that physical intervention has been widely used to maintain range of motion in the paretic upper extremity function post-stroke.

Bilateral training suggests that movement of the non-paretic limb positively affects that of the paretic limb when performed together [18]. This can take the form of swinging the lower extremity while at a sitting position or raising both arms simultaneously depending on whether the lower or upper extremity is affected or both. This can be undertaken in both a symmetric and asymmetric manner and within the home-based environment [22].

### 3.3.3 Motor Skill Learning

Following stroke, part of the sufferer's neural system controlling movement may suffer significant damage. To regain the motor skills, neurorehabilitation is essential [23]. Constraint-induced movement therapy (CIMT), motor imagery and biofeedback are some of the interventions in this category. CIMT involves restraining the unaffected limb to allow for the use of the impaired limb. It is a multifaceted rehabilitation technique which can be easily performed and monitored within the home environment [24]. Delden et al. [22] highlighted the effectiveness of the CIMT in

upper extremity dysfunction correction post-stroke [25–27]. A further study on augmented exercise therapy by Kwakkel et al. [19] also suggested that a 16-h treatment for the first 6 months post-stroke can significantly improve a participant’s motor recovery.

### ***3.3.4 Mirror Imagery and Mirror Neurons Intervention***

The major components of these intervention methods include movement learning, observation and imitation; mirror therapy and mental practice with motor imagery [18].

Mirror imagery implies having a mirror reflection of the non-paretic limb as though it were the affected one [28]. Carvalho et al. [29] stressed that advances in this approach could bring an innovative rational therapeutic intervention to post-stroke sufferers. The main advantages of the motor imagery approach include ease of use, cost and home-based application.

Likewise, the mirror neurons concept involves cognitive coding, movement observation and imitation to help restore motor function post-stroke. The concept hypothesized that performing a goal-directed motor function can activate sympathetic motor neurons in the observer. Nevertheless, there was no significant evidence to support this hypothesis in helping motor function recovery post-stroke [14, 18].

### ***3.3.5 Adjuvant Therapies***

Adjuvant therapies include all the prescribed stimulation and assistive drugs which enhance the recovery of post-stroke sufferers. These include electrical stimulation of the affected limb, invasive and non-invasive stimulation of the patient’s brain and drug stimulation [18]. Furthermore, some of these therapies are also classified as technology-based rehabilitation and are mentioned for significant improvement to post-stroke sufferers [18]. Moreover, they can be combined with home-based remedies for significant outcomes.

### ***3.3.6 Technology-Based Rehabilitation***

Most technology-based neurorehabilitation for post-stroke sufferers work in combination with the previously described physical therapies. These are sometimes regarded as technology-supported training (TST) such as games, virtual reality (VR), robotic-assisted devices and music support [18].

VR and games enable users to interact with a virtual world. The advanced VR interventions help users to experience a computer-generated 3D environment.



Furthermore, task-driven training, which simulates a real-world experience and motivates users to retrain an affected limb, are also included in these interventions [18].

From the aforementioned intervention methods, it is noteworthy that many of them such as the multiple exercises, isolated approaches and motor skill learning, amongst others, can be performed and monitored within the home-based settings, thereby reducing a considerable amount of burden on rehabilitation facilities and logistical constraints like transportation and appointment schedules on the part of post-stroke sufferers. In addition, problems associated with the use of wearable devices such as battery life, which requires the user to frequently remember to charge and wear these devices, can be avoided by embracing home-based unobtrusive sensing solutions. In view of these challenges posed by wearable devices and rehabilitation facilities, this Chapter proposes a novel non-obtrusive monitoring system for post-stroke rehabilitation intervention based on Doppler radar and thermal sensing solutions. This system is intended to deliver feedback on the performance of prescribed post-stroke rehabilitation exercises within a home environment.

### 3.4 Unobtrusive Sensing Solutions

A sensor is a device that can acquire data from another device or system by either direct or non-obtrusive means with the help of changes in its electrical, chemical, magnetic, biological or thermal properties. Common examples of sensors include radar, pressure, ultrasound, motion, force, light (optical), humidity and thermal [30].

#### 3.4.1 *Unobtrusive Thermal Sensing Technology*

Thermal sensing technology have been successfully applied in many areas [31, 32]. Thermopile sensors, for example, consist of two dissimilar metals that convert the temperature gradient between two points into a voltage. A phenomenon which appropriates the linear proportionality between voltage and temperature changes in two dissimilar thermocouple materials is termed the Seebeck effect [30].

A recent study by Mukherjee and Saha [33] on precision thermocouple amplifier further explored the Seebeck principle in finding solutions to temperature monitoring with thermocouples for improved reliability and accuracy. This included using an infrared thermocouple-powered application to measure infrared wavelengths emitted by a human body. Furthermore, the relationship between radiation energy emission and temperature was used to unobtrusively calculate the surface temperature of the body [34]. Non-obtrusive thermal sensing solutions can be applied in both indoor and outdoor environments [35–38]. A review by Sobrino et al. [34] identified more than 30 applications of infrared sensors in different fields such as surveillance and security; building and construction; hazards and health.

### ***3.4.2 Unobtrusive Radar Sensing Technology***

Radar sensing platforms have improved in recent years due to advances in semiconductor fabrication. These improvements are demonstrated in areas such as miniaturization of radar technology, increased radar sensing precision: up to millimetre wave frequencies [39], advancements in air defence systems, vehicle anti-collision designs, meteorological monitoring, guided missile systems and remote sensing systems. Li et al. [40] reviewed recent advances in portable radar systems for motion detection, imaging and ranging to include high sensitivity, high signal-to-noise ratio and improved signal processing systems.

Signal processing is a key component of radar technology. In pulse Doppler radar, for instance, signal processing is performed in a logical and consistent manner which implies taking up many pulses and measuring the phase-shifts across them. This process helps to determine the velocity and range of its targets. In addition, mobile targets can be easily detected by the radar sensor either by the Doppler effect or fast Fourier transforms used in radar signal processing algorithms [41].

### ***3.4.3 Unobtrusive Optical Sensing Technology***

An optical sensor is a device that can change light energy into electrical signals. It has been the subject of many studies in recent years [2]. Common types of optical devices are lasers and fibres. Some properties of optical sensors are wide dynamic range, negligible electromagnetic interference, high sensitivity, electrical isolation, flexible configuration and multiplexing capabilities [2].

Furthermore, optical sensors can be applied in areas such as data transmission, process control, medical imaging and metrology, amongst others [2]. As an example, in structural engineering, fibre optic sensors are used to ascertain the length and propagation speed of concretes in concretes setting analysis [2].

### ***3.4.4 Unobtrusive Ultrasound Sensing Technology***

An ultrasonic sensor is a device which measures the distance between itself and an object with the help of sound waves. In ultrasonic sensing, sound energy is transmitted to and from a target at a specific frequency [42]. It consists of a transmitter and receiver operating at the same frequency. It also has a transducer (driven by a multivibrator) which receives an input from two NAND gates connected in an inverted logic arrangement. After a signal has encountered a target and is reflected back, the transducer converts the signal into an ultrasonic vibration. Some examples of ultrasonic sensors include proximity sensors and switches, retro-reflective and through-beam sensors [42].

The time taken for the signal to hit its target and return to the ultrasonic receiver is known as time of flight (ToF). Figure 3.3 illustrates the ToF principle with transmitted and reflected signals [43].

The TOF principle is very useful in distance calculation of targets in non-obtrusive sensors. Its application cuts across sensor types ranging from ultrasonic to radar and optical sensors. Table 3.1 presents a comparative tabulation of sensor types and their properties.

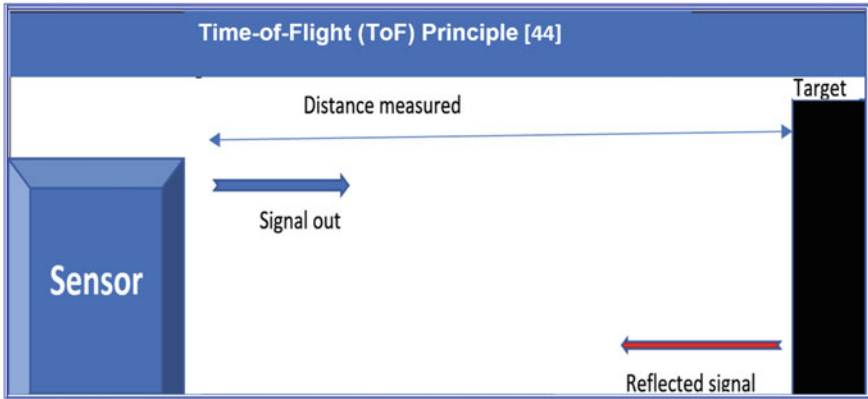


Fig. 3.3 Time of flight principle showing transmitted and reflected signals

Table 3.1 Comparison of low-cost unobtrusive sensors in healthcare applications

| Properties   | Type        |             |             |             |             |
|--------------|-------------|-------------|-------------|-------------|-------------|
|              | Ultrasonic  | Radar       | Optics      | Thermal     | Depth       |
| Precision    | High        | High        | High        | Medium      | Medium      |
| Efficiency   | Medium      | High        | High        | High        | High        |
| Resolution   | Medium      | High        | High        | Medium      | Low         |
| Temperature  | Dependent   | Independent | Independent | Independent | Independent |
| Illumination | Independent | Independent | Dependent   | Independent | Independent |
| Interference | Yes         | No          | No          | No          | Yes         |
| Privacy      | High        | High        | Low         | High        | Medium      |
| Range        | Low         | High        | High        | Medium      | Low         |

## 3.5 Sensor Data Fusion in Unobtrusive Sensing Technology and Applications

### 3.5.1 Sensors Data Fusion

Sensor data fusion (SDF) is a combination of data sets from two or more similar (homogeneous) sensors or dissimilar (heterogeneous) sensors to produce a complementary, cooperative or competitive sensing solution [44]. Many processes are involved in the data set combination. These include data integration, aggregation, pre-processing, filtering, estimation and prediction [44]. The procedures for a SDF arrangement depend on the underlying fusion architecture.

#### 3.5.1.1 SDF Architectures

There are three main types of SDF architectures. These are centralized, distributed and hybrid architectures. The centralized architecture is often applied in homogeneous sensor arrangements. It involves time-based synchronization, correction and transformation of all raw sensor data for central processing. Others include alignment, gating and association as presented in Fig. 3.4 [45]. Furthermore, a central track management (CTM) algorithm keeps track of updates in its target's disposition. CTM is easily implemented in a centralized architecture due to the availability of all raw data in the central processing unit. Filtering and output prediction succeed the CMT algorithm (Fig. 3.4) [45].

In a distributed SDF architecture, data processing for each sensor takes place separately but fusion. This implies that, unlike the centralized architecture, gating, association, local track management, filtration and prediction are done locally for each sensor before fusion of the local tracks shown in Fig. 3.5 [45]. This architecture mainly fits heterogeneous sensors with dissimilar data frames such as infrared and

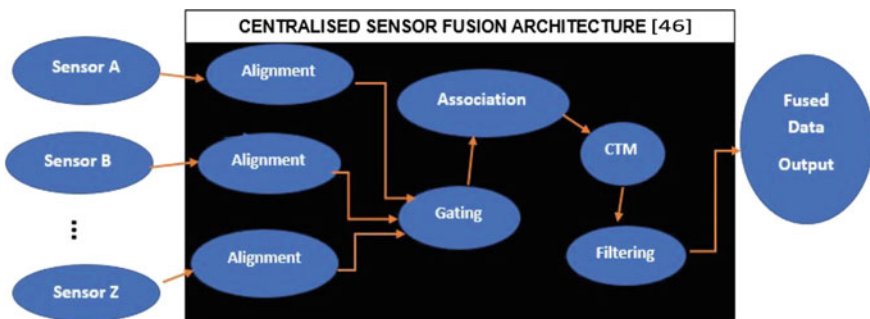


Fig. 3.4 Centralized SDF architecture

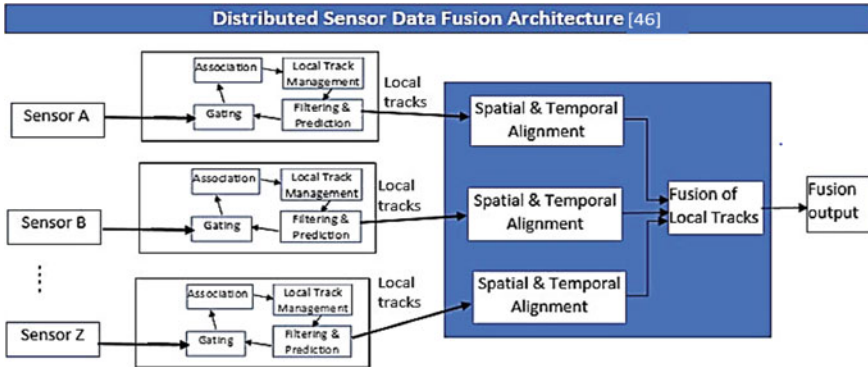


Fig. 3.5 Distributed SDF architecture

radar sensors. Furthermore, data filtering for each sensor is performed by Kalman filtering (KF), square root information or extended KF [46].

The hybrid SDF architecture involves the unification of both centralized and distributed architectures depending on the computational workload, communication or accuracy requirements. In addition, the hybrid SDF provides the merits of a centralized architecture, which include accurate data association and tracking and direct logic implementation. Nevertheless, it is complex and requires high data transfer between the central and local trackers compared with either the centralized or distributed architectures [45].

### 3.5.2 Sensor Data Fusion Applications

Kim et al. [47] proposed a radar and IR sensor fusion system for object detection based on a radar ranging concept which required the use of a calibrated infrared camera alongside Levenberg–Marquardt optimization method. The experiment involved a car with a magenta and a green cross marks as calibrated points (in meters) at different distances. The performance of this experiment was rated 13 times better compared with baseline methods [47].

Furthermore, a double weight-based SAR and infrared sensor fusion research [48] presented a method for automatic ground target recognition (ATR) using a deep learning algorithm. Three levels of fusion were implemented in the algorithm. The pixel-level fusion scheme was used in homogeneous sensor settings [49]. In feature-level fusion, data features were extracted from each sensor automatically using multiple kernel learning-based algorithms while decision-level fusion isolated individual classifier data for output prediction [48].

Another study by Puente et al. [50] discussed the fusion of ground penetrating radar (GPR), infrared thermography (IRT) and terrestrial light detection and ranging (T-LiDAR) sensors. The latter was used for high-resolution models in 3D [50]. Guan

et al. [49] considered the fusion of optical, thermal, fluorescence and microwave satellite data for large-scale crop yields estimation [49].

### **3.6 Unobtrusive Sensing Solutions in Healthcare**

Unobtrusive sensing solutions have gained interest in healthcare management and rehabilitation applications in recent years. These have included: ‘detection of human emotions: blue eyes technology’ [51]; ‘development and evaluation of a non-obtrusive patient monitoring system with smart patient bed’ [52]; recognition of activities of daily living (ADL) for lonely elderly adults through unobtrusive sensing [53]; ‘unobtrusive monitoring to detect depression for the elderly with chronic illness’ [54] and ‘unobtrusive sensing and wearable device for soldiers using WSN’ [55] to mention but a few.

#### ***3.6.1 Health Monitoring with Bio-sensory System***

The use of bio-sensory systems to monitor the health conditions of soldiers to know if they are dead or alive in war times is another example of a combination of unobtrusive and wearable sensing solutions. This is achieved with the help of the global positioning system (GPS) and the GSM networks which provide for the longitudinal and latitudinal information and location of a tracked soldier [55]. Nevertheless, other sensing solutions which are completely unobtrusive, without any wearable cuffs or entanglements, are discussed in the following subsections.

#### ***3.6.2 Monitoring of Vital Signs***

The intelligent phased-array sensor, which has no wires, electrodes nor wearable attachments, is used to remotely monitor vital signs like the heart and respiratory rates in patients. This Doppler-based system uses radiofrequency signals for non-contact vital signs (NCVS) data acquisition on a long-term basis. The process of data acquisition by the sensor to include the propagation of transmitted (TX) and received (RX) signals to and from the target [56].

#### ***3.6.3 Detection of Anomalies in Activities of Daily Living***

Diraco et al. [57] discussed home-based monitoring of elderly patients using radar-based smart sensors to detect abnormalities in ADLs based on the information

provided by an ultra-wideband radar sensor on movement and the cardiorespiratory functions of the body irrespective of physical obstruction and lighting conditions. Experimental results indicated that the smart sensor could distinguish between ADLs and dangerous activities. In addition, unsupervised learning and fall detection using micromotion signatures were also recorded with more than 90% accuracy [57].

Furthermore, human respiratory and motor functions were recorded remotely using passive radar monitoring applications [58]. The data provided a good reference for the health condition of the technology users in-line with their physical activities. For babies, records of their sleep patterns overnight helped breathing monitoring for the prevention of obstructive sleep apnoea (OSA) and sudden infant death syndrome (SIDS) [58].

Nevertheless, Haux et al. [8] reviewed the past, present and future of health-enabling and ambient assistive technologies using original articles in medical informatics and independent survey of key projects. Findings from the study identified battery issues and frequent removal of wearable sensors by users as some of the problems to be solved in future healthcare technological interventions [8]. This review further suggested the need for non-obtrusive sensing solutions for home-based post-stroke rehabilitation that were neither battery-powered nor wearable.

### **3.7 Unobtrusive Sensing for Home-Based Post-stroke Rehabilitation Exercises Using Heterogeneous Sensors: A Case Study**

Information gathered with the help of two or more heterogeneous sensors can be combined for effective monitoring of rehabilitation exercises post-stroke in a home environment. An example is the fusion of data from Doppler radar and thermal sensors. These two sensors are considered suitable due to their added advantages and characteristics (refer to Table 3.1). Some of these include that Doppler radar does not interfere with other frequencies and legacy systems. Thermal sensor, on the other hand, protects the privacy and estimates the postures of users.

#### ***3.7.1 The Proposed Monitoring Approach***

These sensing solutions explore the body temperature and the Doppler shift effect propounded by Christian Doppler in 1842 to estimate the rate of displacement (velocity) of a user's upper extremity function relative to the body whilst in standing or sitting positions. The Doppler principle can be applied to a source at rest or in motion [59].

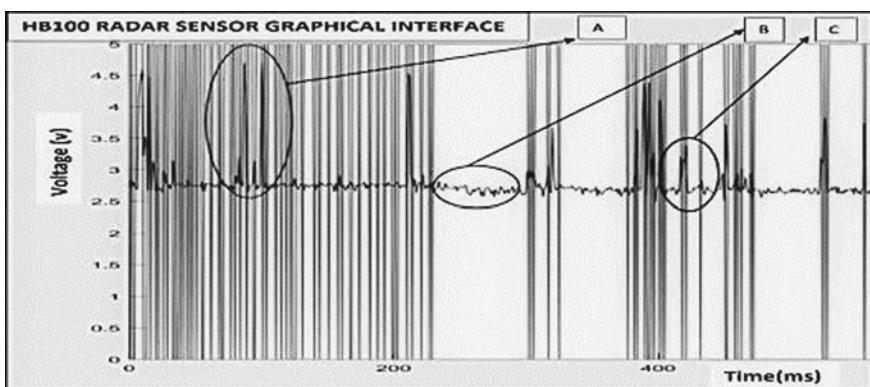
### 3.7.2 Materials and Methods

An HB 100 Doppler radar (HBDR) sensor and a Heimann HTPA 32×32 (HH32T) thermal sensor was used for this experiment. Each was placed at a fixed position. During the experiments, participants performed upper extremity extension and flexion facing the direction of the sensor in a sitting position while preventing the movement of other parts of the body during the exercises. The HBDR was connected to an Arduino Uno microcontroller and was programmed to run on MATLAB R2018b software. Data from the HH32T were aggregated using *SensorCentral* [60]. Raw data from both sensors were collected, processed and filtered and their outcome were considered separately.

### 3.7.3 Experimental Results and Discussion

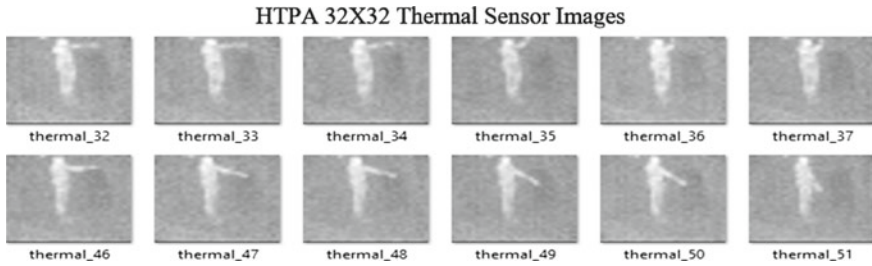
The Doppler sensor generated two signal types as presented in Fig. 3.6. These were digital and analog. The latter operated from 2.5 to 5 V indicating changes in voltage levels with respect to the displacement of the upper limb at instances of flexion and extension. These levels are indicated as A, B and C for ‘high’, ‘No’ and ‘low’ displacements, respectively. The digital signal output, however, only indicated 0 V for ‘no displacement’ and 5 V for all other displacement levels (Fig. 3.6).

Thermal images gleaned from the HH32T sensor (Fig. 3.7) showed frames on instances of extension and flexion in all lighting conditions. Nevertheless, it could not compute the rate of displacement (velocity) and the total counts which were required to ensure that the exercises were performed as prescribed; hence, the radar sensor complemented these shortcomings using the Doppler effect.



**Fig. 3.6** HB 100 Radar sensor data. A—High displacement, B—no displacement and C—low displacement





**Fig. 3.7** HTPA 32X32 thermal sensor images during flexion and extension

Furthermore, Table 3.2 illustrates the Doppler sensor behaviour with corresponding thermal sensor images. It could be observed in Fig. 3.7 that on thermal\_32 that when the upper limb was at full extension, the Doppler sensor (refer to Fig. 3.6) indicated a high displacement (A). These actions were also observed on thermal\_35 (flexion), thermal\_47 (low extension) and thermal\_51 (little/no extension), to have matched with instances high displacement, low displacement and little/no displacement respectively, amongst others.

The second iteration of this work will involve the implementation of the distributive fusion architecture and development of a user-friendly interface for feedback to users on their completion of prescribed exercises.

**Table 3.2** Thermal sensor frames and corresponding radar sensor behaviour

| S/N | Thermal frame | Thermal sensor       | Radar sensor           |
|-----|---------------|----------------------|------------------------|
| 1   | Thermal_32    | Full extension       | High displacement      |
| 2   | Thermal_33    | Full extension       | High displacement      |
| 3   | Thermal_34    | Flexion              | High displacement      |
| 4   | Thermal_35    | Flexion              | High displacement      |
| 5   | Thermal_36    | Flexion              | High displacement      |
| 6   | Thermal_37    | Flexion              | High displacement      |
| 7   | Thermal_46    | Full extension       | High displacement      |
| 8   | Thermal_47    | Low extension        | Low displacement       |
| 9   | Thermal_48    | Low extension        | Low displacement       |
| 10  | Thermal_49    | Low extension        | Low displacement       |
| 11  | Thermal_50    | Low extension        | Low displacement       |
| 12  | Thermal_51    | Little/ no extension | Little/no displacement |

### 3.8 Conclusion and Future Work

This Chapter explained the concept of stroke and a range of rehabilitation techniques aimed at assisting post-stroke sufferers recover from neuromuscular dysfunctions. While wearable devices were noted to have a number of shortcomings such as wearability and adoption, the rehabilitation facilities were found to have logistical complexities. These challenges amongst others informed the proposed approach based on unobtrusive (radar and thermal) sensing solutions. Initial experimental results presented the data of human participants at instances of flexion and extension. Also, the radar sensor indicated different displacement levels of the upper limb.

Future work will attempt to incorporate the distributive sensor fusion algorithms within the overall system. It will also seek to estimate the angular displacement of the upper limb with reference to global coordinates to ensure that the paretic limb is properly stretched during the exercises. The final components of the system will include an avatar user interface development to provide an interactive feedback to users to guide them in the completion of their exercises.

**Acknowledgements** This project is supported by the European Union's INTERREG VA Programme, managed by the Special EU Programmes Body (SEUPB).

### References

1. Wilson CB (1999) Sensors in medicine 319:13–15
2. Dhiraj A, Deepa P (2012) Sensors and their applications. *J Phys E: Sci Instrum* 1(5):60–68
3. Spring S, Sutton WM (2015) Classification overview
4. Lymberis A (2000) Smart wearables for remote health monitoring, from prevention to rehabilitation: current R&D, future challenges. In: Proceedings of the 4th annual IEEE conference on information technology applications in biomedicine, UK, vol 7, no 1, pp 25–4888
5. Giorgino T, Tormene P, Maggioni G, Pistarini C, Quaglini S (2005) Wearable kinesthetic system for capturing and classifying upper limb gesture in post-stroke rehabilitation. *J NeuroEng Rehabil* 2:1–16
6. Whipple RH (1970) Specificity of speed of exercise *T* 1692–1700
7. Fung J, Richards CL, Malouin F, McFadyen BJ, Lamontagne A (2006) A treadmill and motion coupled virtual reality system for gait training post-stroke. *CyberPsychol Behav* 9(2):157–162
8. Haux R, Koch S, Lovell NH, Marschollek M, Nakashima N, Wolf K-H (2016) Health-enabling and ambient assistive technologies: past, present, future. *Yearb Med Inf* 25(S 01):S76–S91
9. D'Aliberti G, Longoni M, Motto C, Oppo V, Perini V, Valvassori L, Vidale S (2017) Emergency management in neurology, pp 1–91
10. Stroke Association (2018) State of the nation stroke statistics, February
11. "Pathophysiology\_Neuro4Students," Neuro4Students, 2010. [Online]. Available: <https://neuro4students.wordpress.com/pathophysiology/>
12. Morgenstern LB, Hemphill JC III, Anderson C, Becker K, Broderick JP, Connolly ES Jr, Greenberg SM, Huang JN, Macdonald RL, Messé SR, Mitchell PH, Selim M, Tamargo RJ (2010) Guidelines for the management of spontaneous intracerebral hemorrhage. *Stroke* 41(9):2108–2129
13. Wieroniy A (2016) A new era for stroke. *Br J Neurosci Nurs* 12:S6–S8

14. Raffin E, Hummel FC (2018) Restoring motor functions after stroke: multiple approaches and opportunities. *Neuroscientist* 24(4):400–416
15. Kalra L, Ratan R (2007) Recent advances in stroke rehabilitation 2006. *Stroke* 38(2):235–237
16. Cramer SC (2008) Repairing the human brain after stroke: I. Mechanisms of spontaneous recovery. *Ann Neurol* 63(3):272–287
17. Chollet F, Dipiero V, Wise RJS, Brooks DJ, Dolan RJ, Frackowiak RSJ (2018) The functional anatomy of motor recovery after stroke in humans: a study with positron emission tomography. *Ann Neurol* 29(1):63–71
18. Hatem S, Saussez G, della Faille M, Prist V, Zhang X, Dispa D, Bleyenheuft Y (2016) Rehabilitation of motor function after stroke: a multiple systematic review focused on techniques to stimulate upper extremity recovery. *Front Hum Neurosci* 10:bl 442
19. Kwakkel G, Peppe R, Wagenaar R, Dauphinee C, Richards S, Ashburn A, Kimberly M, Lincoln N, Partridge C, Wellwood I, Langhorne P (2004) Effects of augmented exercise therapy time after stroke: a meta-analysis. *Stroke* 35(11):2529–2536
20. Wang R, Chen H, Chen C, Yang Y (2005) Efficacy of Bobath versus orthopaedic approach on impairment and function at different motor recovery stages after stroke: a randomized controlled study. *Clin Rehabil* 19(2):155–164
21. Pollock A, Baer G, Campbell P, Choo P, Forster A, Morris J, Pomeroy V (2014) Physical rehabilitation approaches for the recovery of function and mobility following stroke
22. Delden A, Peper C, Harlaar J, Daffershofer A, Zipp N, Nienhuys K, Koppe P, Kwakkel G, Beek P (2009) Comparing unilateral and bilateral upper limb training: the ULTRA-stroke program design. *BMC Neurol* 9(1):1–14
23. Muratori LM, Lamberg EM, Quinn L, Duff SV (2013) Applying principles of motor learning and control to upper extremity rehabilitation. *J Hand Ther Off J Am Soc Hand Ther* 26(2):94–102, quiz 103
24. Sirtori V, Corbetta D, Moja L, Gatti R (2009) Constraint-induced movement therapy for upper extremities in stroke patients (review). *Cochrane Rev* (4):4–6
25. Wolf SL, Winstein CJ, Miller JP (2006) Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after stroke: the excite randomized clinical trial. *JAMA* 296(17):2095–2104
26. Langhorne P, Coupar F, Pollock A (2009) Motor recovery after stroke: a systematic review. *Lancet Neurol* 8(8):741–754
27. Page SJ, Levine P, Leonard A, Szaflarski JP, Kissela BM (2008) Modified constraint-induced therapy in chronic stroke: results of a single-blinded randomized controlled trial. *Phys Ther* 88(3):333–340
28. Wittkopf PG, Johnson MI (2017) Mirror therapy: a potential intervention for pain management. *Rev Assoc Méd Bras* 63(11):1000–1005
29. Carvalho D, Teixeira S, Lucas M, Yuan T-F, Chaves F, Peressutti C, Machado S, Bittencourt J, Menéndez-González M, Nardi AE, Velasques B, Cagy M, Piedade R, Ribeiro P, Arias-Carrión O (2013) The mirror neuron system in post-stroke rehabilitation. *Int Arch Med* 6:41
30. Huynh T (2015) Fundamentals of thermal sensors
31. Hevesi P, Wille S, Pirkel G, Wehn N, Lukowicz P (2014) Monitoring household activities and user location with a cheap, unobtrusive thermal sensor array. In: *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing—UbiComp’14 Adjunct*, pp 141–145
32. Trofimova AA, Masciadri A, Veronese F, Salice F (2017) Indoor human detection based on thermal array sensor data and adaptive background estimation, March
33. Mukherjee A, Saha T (2018) Precision Thermocouple Amplifier for substrate temperature monitoring in an ECR-PE nano-film deposition system. In: *IEEE*, vol 4, no 18
34. Sobrino JA, Del Frate F, Drusch M, Jiménez-Muñoz JC, Manunta P, Regan A (2016) Review of thermal infrared applications and requirements for future high-resolution sensors. *IEEE Trans Geosci Remote Sens* 54(5):2963–2972
35. Berni JAJ, Zarco-Tejada PJ, Suárez L, González-Dugo V, Fereres E (2009) Remote sensing of vegetation from UAV platforms using lightweight multispectral and thermal imaging sensors. *Int Arch Photogramm Remote Sens Spatial Inform Sci* 38:6 pp

36. Dinh T, Phan H, Qamar A, Woodfield P, Nguyen N, Dao DV (2017) Thermoresistive effect for advanced thermal sensors: fundamentals, design considerations, and applications 26(5):966–986
37. Khanal S, Fulton J, Shearer S (2017) An overview of current and potential applications of thermal remote sensing in precision agriculture. *Comput Electron Agric* 139:22–32
38. Inman K, Wang X, Sangeorzan B (2010) Design of an optical thermal sensor for proton exchange membrane fuel cell temperature measurement using phosphor thermometry. *J Power Sour* 195(15):4753–4757
39. Al-Hourani A, Evans R, Farrell P, Moran B, Martorella M, Kandeepan S, Skafidas S (2018) Millimeter-wave integrated radar systems and techniques. In: *Academic press library in signal processing*, vol 7. Elsevier, pp 317–363
40. Li C, Peng Z, Huang TY, Fan T, Wang FK, Horng TS, Munoz-Ferreras JM, Gomez-Garcia R, Ran L, Lin JA (2017) A review on recent progress of portable short-range noncontact microwave radar systems. *IEEE Trans Microw Theory Tech* 65(5):1692–1706
41. Parker M, Parker M (2017) Pulse doppler radar. In: *Digital signal processing*, vol 101. Elsevier, pp 241–251
42. Fidanboyly KA, Efendioglu HS (2011) Fiber optic sensors and their applications. In: *Symposium a quarterly journal in modern foreign literatures*, May, pp 1–6
43. Bilal MR (2017) Ultrasonic sensor working applications and advantages. *Microcontrollers Lab*
44. Terabee (2016) Time-of-flight principle, measuring the distance between sensor/object
45. Jitendra R (2013) Multi-sensor data fusion with MATLAB, vol 106, no 11. CRC Press, 6000 Broken Sound Parkway NW
46. Lytrivis P, Thomaidis G, Amditis A, Lytrivis P, Thomaidis G (2009) Sensor data fusion in automotive applications. *INTECH*, February, p 490
47. Kim T, Kim S, Lee E, Park M (2017) Comparative analysis of RADAR-IR sensor fusion methods for object detection. In: *ICCAS*, pp 1576–1580
48. Kim S, Song W, Kim S (2018) Double weight-based SAR and infrared sensor fusion for automatic ground target recognition with deep learning. *Remote Sensing* 72(10): 2072–4292
49. Guan K, Wu J, Kimball J, Anderson M, Froelking S, Li B, Hain C, Lobell D The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sens of Environ* 199:333–349
50. Ivan P (2018) Reconstructing the Roman Site ‘Aquis Querquennis’ 379(10):1–16
51. Sondhi A (2017) Detecting human emotions: blue eyes technology 2(6):12–16
52. Van Dijk R, Liang W, Zhang B (2017) Development and evaluation of a non-obtrusive patient monitoring system with smart patient beds. In: *Distributed, ambient and pervasive interactions*, pp 482–490
53. Meng L, Miao C, Miao C, Leung C (2017) Towards online and personalized daily activity recognition, habit modeling, and anomaly detection for the solitary elderly through unobtrusive sensing. *Multimedia Tools Appl* 76(8):10779–10799
54. Kim JY, Liu N, Tan HX, Chu CH (2017) Unobtrusive monitoring to detect depression for elderly with chronic illnesses. *IEEE Sens J* 17(17):5694–5704
55. Kumar MS (2017) Unobtrusive sensing and wearable device for soldiers using WNS 3(10):580–587
56. Hall T, Lie DYC, Nguyen TQ, Mayeda JC, Lie PE, Lopez J, Banister RE (2017) Non-contact sensor for long-term continuous vital signs monitoring: a review on intelligent phased-array doppler sensor design. *Sensors (Switzerland)* 17(11):1–20
57. Diraco G, Leone A, Siciliano P (2017) A radar-based smart sensor for unobtrusive elderly monitoring in ambient assisted living applications. *Biosensors* 7(4)
58. Li W, Tan B, Piechocki R (2018) Passive radar for opportunistic monitoring in e-health applications. *IEEE J Transl Eng Health Med* 6, September 2017
59. TutorVista.com (2018) Doppler shift formula 7440(iv):1–4
60. Rafferty J, Synnott J, Nugent C, Ennis A, Catherwood P, McChesney I, Cleland I, McClean S (2018) A scalable, research oriented, generic, sensor data platform. *IEEE Access* 6:bl1 45473–45484

# Chapter 4

## Lessons from Hands-Free Data Entry in Flexible Cystoscopy with Glass for Future Smart Assistance



Charles Templeman, Francisco Javier Ordoñez Morales, Mathias Ciliberto, Andrew Symes and Daniel Roggen

**Abstract** We explore how Google Glass can be used to annotate cystoscopy findings in a hands-free and reproducible manner by surgeons during operations in the sterile environment inspired by the current practice of hand-drawn sketches. We present three data entry variants involving head movements and speech input. In an experiment with eight surgeons and foundation doctors having up to 30 years' of cystoscopy experience at a UK hospital, we assessed the feasibility, benefits and drawbacks of the system. We report data entry speed and error rate of input modalities and contrast it with the participants' feedback on their perception of usability, acceptance and suitability for deployment. These results offer an expanded analysis of the participants' feedback compared to previous analysis. The results highlight the potential of new data entry technologies and point out directions for future improvement of eyewear computers. The findings can be generalised to other endoscopic procedures (e.g. OGD/laryngoscopy) and could be included within hospital IT in future. The source code of the Glass application is available at <https://github.com/sussexwearlab/GlassMedicalDataEntry>.

**Keywords** Eyewear computer · Gestural interface · Speech interface, HCI · Data entry

### 4.1 Introduction

Assistance in smart homes needs not only to rely on sensors deployed in the environment: it can also make use of mobile and wearable devices for sensing conditions, interacting with the user or inferring his or her context. Eyewear computers are

---

C. Templeman  
Brighton and Sussex Medical School, Brighton, UK

C. Templeman · F. J. Ordoñez Morales · M. Ciliberto · D. Roggen (✉)  
University of Sussex, Brighton, UK  
e-mail: [d.roggen@sussex.ac.uk](mailto:d.roggen@sussex.ac.uk)

A. Symes  
Brighton and Sussex University Hospital, Brighton, UK

© Springer Nature Switzerland AG 2020

F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications and Networks, [https://doi.org/10.1007/978-3-030-25590-9\\_4](https://doi.org/10.1007/978-3-030-25590-9_4)

particularly interesting in smart assistance scenarios, as more than 75% of the population will wear prescription glasses above the age of 50–64-years old, and this number increases to more than 83% of the population for those aged 65 and above. In other words, glasses equipped with sensors, computing capabilities and interaction modalities would form an ideal platform to provide smart assistance for the elderly. Google Glass is a particularly interesting platform for this purpose, as it was one of the first eyewear computers released with a complete software development kit, a myriad of sensors (e.g. front-facing camera, microphone, inertial measurement unit and eye-blink sensor) and multimodal interaction modalities with visual and auditory overlay.

The successful design of assistive devices requires a combination of clear benefits for the user, ease of use, comfort, handling potential stigma associated with being seen with such a device, affordable costs, and others.

In this work, we have investigated whether Google Glass can be a suitable platform to provide smart assistance in the particular clinical context of data entry in cystoscopy. The clinical context has often much more stringent constraints than home assistance, such as rapidity of use, precision, suitability for the sterile environment, and others. In doing the study, we present in this chapter, and we believe a number of lessons can be learned for future smart home assistance with wearable computers.

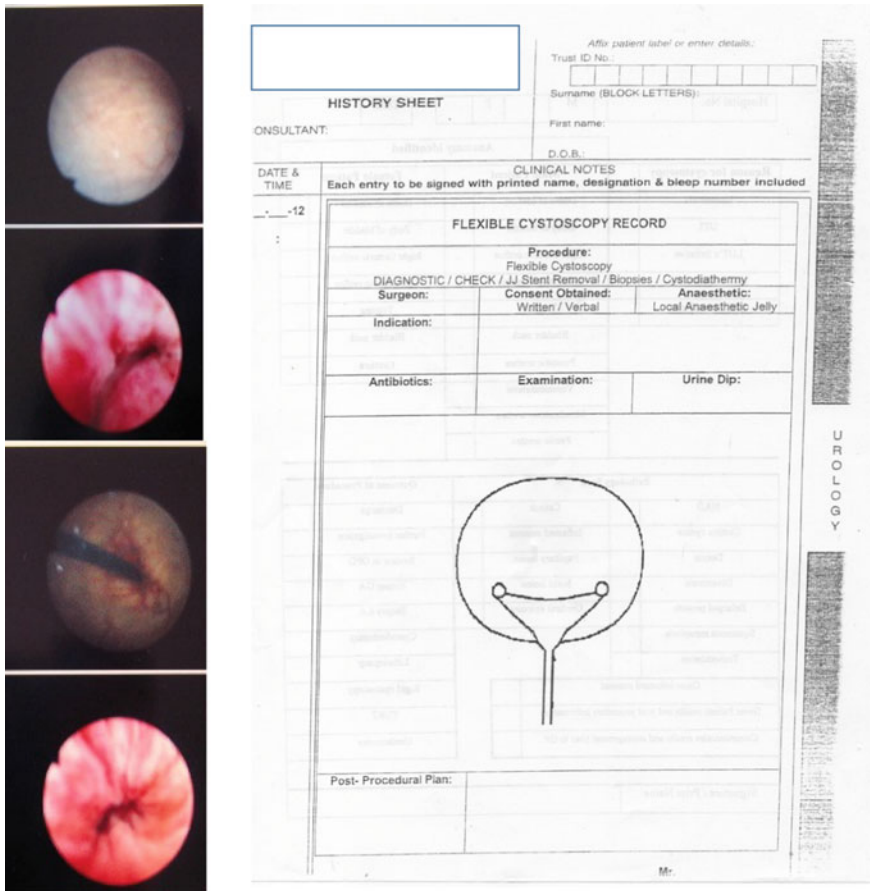
With an estimated 151,000 new cases diagnosed in Europe in 2012 and 429,000 new cases worldwide in 2012, bladder cancer is one of the most common cancers among both men and women [1]. In the UK, despite bladder cancer rates dropping by between 25 and 35% since 1990, it remains the seventh most common cancer-related cause of death, reflecting both the aggressive nature of the disease and the inherent difficulties in diagnosis and management. As cigarette smoking is the major risk factor, an increase in incidence in the developing world is expected in future [2, 3].

The majority of bladder cancer (~80%) is known as non-muscle invading bladder cancer (NMIBC) and generally presents with painless symptoms, either haematuria (blood in the urine which is not always visible) or urinary symptoms easily attributed to infection such as dysuria (pain while passing urine) [4, 5]. NMIBC ranges from low-grade to high-grade with associated rises in risk of mortality but even slow-growing, low-grade tumours have an extremely high rate of recurrence post-treatment of around 31% rising to 78% for high-grade tumours [2, 6]. These recurrences require multiple follow-up procedures and lengthy surveillance over several years. Late presentation to health care providers, aggressive disease and high recurrence rates combine with technological strategies for diagnosis and management that are expensive and require a high degree of clinical expertise to cause NMIBC to be considered one of the costliest cancers per capita. This is both in terms of financial and human cost; diagnosis, treatment and subsequent long-term surveillance can have a substantial and negative impact on the quality of life [6–8].

The diagnostic gold standard in NMIBC is cystoscopy, the use of a flexible endoscope inserted through the urethra in order to image the urinary tract and bladder as well as perform biopsies to enable grading and staging of cancers found, resection of tumours and introduction of chemotherapeutic agents. This imaging is also used as a reference point for subsequent surgery where necessary. While guidelines issued by

clinical bodies such as the UK's National Institute for Clinical Excellence (NICE) [9] and The Royal College of Surgeons (RCS) [10] provide guidance on performing such procedures and what should be recorded (e.g. size, number and appearance of lesions), there is no real standardised, organisational methodology for how this should be logged to best effect.

We began this project by consulting closely with urological surgeons of various grades and levels of experience to establish where to concentrate our efforts and what improvements would be most useful to them when making patient notes. We found out that static pictures taken from a cystoscope (Fig. 4.1, left) are felt to be insufficient in the documentation of location of tumours due to the lack of reference



**Fig. 4.1** Left: the view through a cystoscope captures only a partial view of the ovoid bladder. Without reference points, it is often not possible to accurately gauge in which direction the cystoscope is oriented: this prevents still images from being used as the sole documentation approach. Right: Example of NHS cystoscopy report template. A hand-drawn schematic bladder diagram is used to indicate location and appearance of lesions

points with which to orient the viewer. Therefore, hand-drawn bladder sketches are often used, either alongside or instead of the captured image (Fig. 4.1, right). The issues identified with this approach were inconsistent reporting methods between users, inaccuracy of drawing from memory and high quantity of paperwork leading to overly complicated patient notes. The key contributions of this chapter are:

- Identification of limitation of current data entry mechanism through interviews with surgeons.
- Development of a hands-free data entry system relying on a comparatively low-cost eyewear computer with three input modalities: speech and discrete or continuous head movements.
- Technical and usability evaluation with eight surgeons and foundation doctors.
- Comparative evaluation of learning curve, data entry speed and error rate of each input modality.
- Qualitative evaluation through exhaustive interviews and questionnaires of the surgeon's perceptions of usability, acceptance and suitability.

## 4.2 State of the Art

### 4.2.1 *Current Cystoscopy Procedure*

Flexible cystoscopy is not routinely performed in a sterile environment but involves the use of sterilised equipment that requires highly trained staff using both hands to operate while at the same time interacting with the patient. In the case of diagnostic imaging, the patient will be left to change into a gown and will then be positioned on a couch in the supine position. The surgeon then hand scrubs using an aseptic technique and dons sterile gloves before handling the previously sterilised cystoscope. From this point, on the surgeon's hands are unavailable for any task outside the operative field as this would require rescrubbing and changing gloves in order to minimise the infection risk.

The surgeon inserts the cystoscope into the urethra before proceeding to the bladder while observing for anatomical landmarks on the display screen. Both hands are required for this task and the imaging procedure typically takes around 10 min. Once finished the cystoscope is withdrawn and the patient can be settled and allowed to redress, taking approximately 5 min. Once the patient has left the room the surgeon is then able to wash once more and only then begins to make notes regarding what has been observed. A surgeon would typically perform several such procedures in a 4/5-h clinic.

Results and findings derived from cystoscopy are recorded after the fact by both written and diagrammatic methods, entailing memory recall on the part of the surgeon. While images and video can be recorded by some cystoscopy equipment, it regarded as impractical by those surgeons we talked to as it is technically difficult to share with other members of multidisciplinary medical teams, and there is no way of



reliably associating the large file sizes necessary with the EHR. Still images can be very useful but due to the structure of the bladder extra notes must still be made by the surgeon in order to clarify the exact position of any abnormalities in relation to anatomical landmarks. The surgeons produce a written description of findings and pair this with a diagrammatic representation of a urinary bladder that is annotated by hand (Fig. 4.1, right). This annotation does not have strictly defined parameters and so can exhibit wide variance between practitioners, although local conventions do dictate to some extent the symbols and the manner in which they are used. The annotated diagram is then kept with the patient's paper medical notes.

A major disadvantage to the current method for gathering data is that of the high number of separate notes and sketches, recorded by multiple clinicians, does lead to a high degree of variance in the reporting of what are objective findings (both in terms of appearance and exact location of lesions). The lack of a persistent annotated image allows for a degree of inconsistency when attempting to track the progression of disease as the same clinician may well not be performing subsequent cystoscopy procedures and individualised style and approaches to notation are common. This can have an impact on the efficiency of analysing public health data as well. Another disadvantage is the reliance on memory recall, inserting an unnecessary burden of potential error into the procedure.

## 4.2.2 *Technical Innovations*

Recent years have seen advances in the provision of high-quality HD still images and video capture to medical practitioners for diagnostic and interventional procedures [11]. These advances often come as a 'stack' comprising surgical instruments and software back-end which require the updating of existing hospital equipment leading to high financial investment (cost can be upwards of £50,000 per 'stack'). However, it appears that annotated notes and still images remain the preferred method of recording such data during flexible cystoscopy, rather than video capture which involves longer examination times [12].

Computerised endoscopy reporting tools exist [13] and go some way towards integrating paper notes and digital images made but offer little more than is already possible as data is still entered after the fact. The current situation is therefore often a combination of digital still images together with handwritten or typed notes and annotated diagrams, either printed out or stored as part of the EHR. Potential novel data entry systems could include foot pedals [14], head-mounted display (HMD) or wearable technology utilising speech to text- or gesture-based input [15–17], gesture-controlled devices [18], incorporating new controls into existing endoscopic equipment [19, 20], or using a 'middleman' to transcribe notes dictated by the clinician which would require additional highly trained personnel. In applying novel technological solutions to these issues, there are multiple points to consider. Foot pedals, although already widely used by surgeons in theatres, are generally used for simple binary actions such as activating tools and even then suffer from issues

of accuracy and poor ergonomics [14, 19, 20]. They would be inappropriate for text input requiring increased time and coordination and an additional screen on which to view the text input, adding a second or third screen to an already cluttered and ergonomically poor environment [21]. Wearable gesture-controlled devices have shown potential in data entry and in controlling external devices, as have ambient gesture sensors, but text input is limited and too slow to be used in these circumstances [22–25]. Adding ‘keyboard’ controls to existing endoscopic equipment could allow input speed to approach normal typing speed [26], but control of the endoscope could be impaired. HMDs and eyewear computers appear to overcome some of these issues. They should not interfere with fine motor control of the endoscope nor have an appreciable impact on infection control. They reduce the ergonomic load of further screens and could combine gesture control and speech input. Google Glass is a much mediatised example of commercial eyewear computers (e.g. Vuzix, Sony SmartEyeGlass). Medical research involving Google Glass and similar devices has concentrated on image capture/video recording capabilities and wireless communication, using these to enhance teaching [27–31], to provide offsite consultations [32–34], to refer to diagnostic imaging or patient data during surgical or interventional procedures [35–37] and to record medical data [32, 38, 39].

Outside of medicine, HMDs and eyewear computers have been investigated for use in on-site inspection utilising both data retrieval and recording [40], for image capture and environmental analysis [41], data recording [42], for capturing and guiding laboratory experiments [43, 44], enhancing teaching [45], guidance systems for manual assembly and repair tasks [15, 46], control of drones [47] and as a replacement for mobile technology for those with physical disability that complicates or impedes use with hands [48].

### 4.3 Methodology and Design Input

Our methodology followed a user-centred design process whereby through a series of discussions with surgical staff aspects of the current procedure that could be improved upon were elicited, using a Glass device and associated promotional material as prompts. Glass is used as a device representative of the capabilities of the eyewear computers. As eyewear computers can be used hands-free, which is a major consideration in clinical environments where infection control is of high importance, we wished to investigate how a device such as Glass could be used to augment current procedures. A key finding was concerned over lack of consistency in reporting the results of cystoscopy due to inter-user variance. This was identified as being a factor of differences in annotation style and results from multiple cystoscopies being recorded separately which can not only make reviewing patient notes complicated but also potentially generates an enormous amount of extraneous paperwork. Through these discussions, it emerged that a diagrammatic representation of the bladder was found to be useful as an adjunct to still images recorded and as a quick reference for consultations but that improvements were possible in the implementation of such a

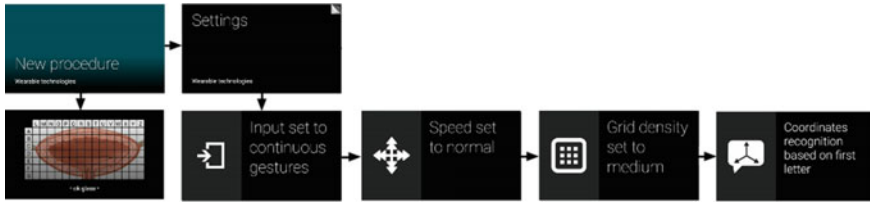
system. As there were also concerns raised regarding the accuracy of the drawings, an idea was put forward that a grid overlaying a basic image of a bladder could be designed together with standardized symbols for notation. To address the issue of consistency, it was decided that a persistent image, annotated in a standardized manner, could be assigned to each patient when first encountered. This image could then be added to during subsequent follow-up, providing a single reference point with a clear notation to aid in planning surgery, possibly reducing the length of time of surgical investigations, and aiding in future medical and surgical management. The choice of input modalities emerged from considerations on the ergonomics of current procedures (see previous section) and the capabilities of Glass.

#### 4.4 Bladder Annotation on Glass

Based on the design input, we developed a digital replica with Glass of the current procedure of annotating a hand-drawn bladder sketch. In the new procedure with Glass, we envisioned that the surgeon would set up Glass prior to inviting the patient to the examination room. In future derivations of the application patient, data and demographics could also be provided to the surgeon at this point. Before entering the surgical field, the surgeon would call up the main UI of the application (Fig. 4.3). All further interaction with Glass would be hands-free to remain within infection control guidelines. Notes akin to those currently made post-procedure could then be made during the procedure itself, negating the requirement to rely on memory. This would allow information to be accurately recorded in a systematic way with direct reference to the view available on the cystoscopy screen. Once the procedure is finished the surgeon would be able to quickly download the Glass file via USB to the same computer used to make notes on the EHR. The image could also be printed for inclusion in the paper notes together with any still images recorded by the cystoscope itself. In terms of the annotation, a prescribed set of symbols could be defined. Together with less reliance on memory, this should increase accuracy and reduce inter-user variability. A persistent bladder diagram could also be saved for subsequent investigations, reducing the amount of paperwork. A persistent image could also make decisions regarding potential surgical interventions simpler as a single image would be necessary to exchange information between physicians, further reducing potential errors.

Therefore, the main UI element is a schematic bladder, where the surgeon can pinpoint a location to annotate with an icon representing the nature of the lesion. The application has been implemented with the Glass Development Kit following an ‘immersion’ design pattern which is a Glass-specific pattern where the application defines its own UI and takes control of the user experience.

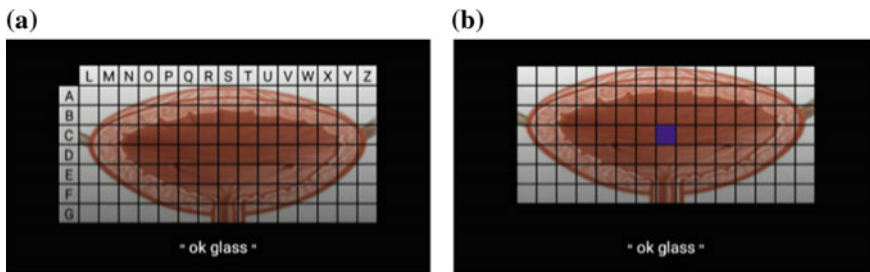
After launching the application, the user is directed to a start screen. From the start screen (Fig. 4.2), it is possible to start a new cystoscopy procedure, which allows surgeons to annotate the bladder image or to move to the settings card to configure the application.



**Fig. 4.2** Application navigation flow. The main timeline comprises two cards: new procedure and settings. The settings are composed by four cards, one per parameter

When a procedure starts a new card is loaded, showing a bladder image. In order to help with data entry, a grid is overlaid on top. It serves as a reference to enter the data observed during the procedure. In order to populate the grid, the user has to choose a cell and indicates the type of annotation to place at that location. The procedure card includes a voice-triggered menu, activated when speaking ‘ok glass’, which contains a list with the voice-activated commands to add or remove annotations. The user then selects the grid cell and finally enters a number from 1 to 5 corresponding to an icon (e.g. a type of lesion) to insert in the cell. The application offers three hands-free modalities with which to choose a grid cell: ‘Speech’, ‘Discrete head movements’ and ‘Continuous head movements’.

In ‘Speech’ mode, the cell is selected by means of two coordinates (row and column) spoken by the user using the NATO phonetic alphabet (Fig. 4.3a). The NATO alphabet represents each of the 26 letters of the English alphabet using phonetically distinguishable code words (e.g. ‘Alpha’, ‘Bravo’). We use the Glass speech recognizer to process the voice input. In speech mode, an additional setting defines how the captured speech is transcribed into coordinates. It can require a ‘Full match’, where the text spoken by the user must exactly correspond to code words in the phonetic alphabet. However, preliminary trials showed that speech recognition errors were frequent. Therefore, the other options are ‘First letter’, where the first letter of the words spoken by the user is matched against the phonetic alphabet and taken as cell



**Fig. 4.3** Depending on the input modality the procedure grid shows: **a** rows and columns identified by indexes for the speech input modality or **b** a cursor highlighting the cell selected for the input mechanisms based on head movements

index, and ‘Similarity’, where the words in the phonetic alphabet which is closer to the words recognized from the user are taken as index. Hereafter we used the ‘First letter’ mode (Fig. 4.3a).

In ‘Discrete movements’ mode, a cursor indicating the cell selected is overlaid to the grid interface (Fig. 4.3b). The user can move the cursor within the grid using head nods vertically and laterally. The gyroscope is employed to calculate the angular velocity of the device. Angular rotation above a threshold indicates a motion event.

In ‘Continuous movements’, the selected cell overlaid on the grid is moved by according to the difference between the current head orientation compared to the initial orientation as measured by the gyroscope. A head rotation of  $\sim 60^\circ$  horizontally and  $\sim 35^\circ$  vertically allows to navigate from one edge to the other on the grid.

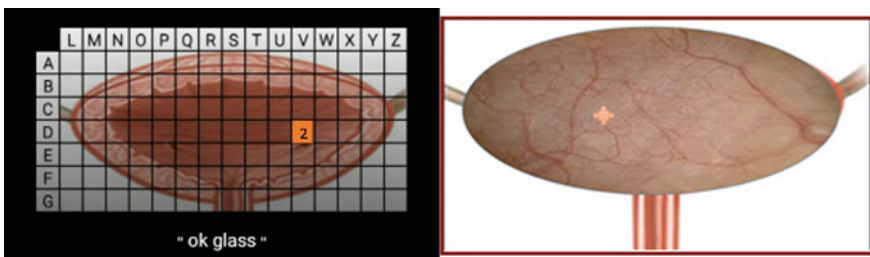
The Glass application continuously recorded how the user interacted with it. Each cystoscopic procedure leads to a log file which contains timestamped information about each interaction with the device (cell movement direction, speech recognizer output, etc.). This allowed us to obtain precise data for subsequent analyses.

The source code of the Glass application is available at <http://github.com/sussexwearlab/GlassMedicalDataEntry>.

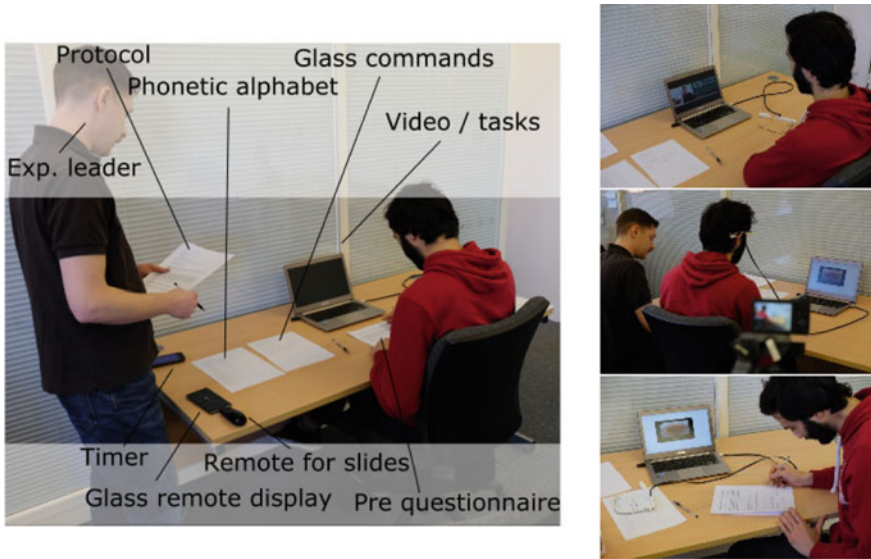
## 4.5 Protocol

The experimental protocol (Fig. 4.5) was designed to collect both quantitative and qualitative data about the suitability of Glass to capture findings in cystoscopy. As for all participants, this would be the first time they would employ Glass and this application, we designed the protocol to assess learning curve, error rate and accuracy of data entry through the objective logs recorded by the application. We set up questionnaires and elicited discussion to compare and contrast these objective findings with the subjective perception of the clinicians regarding usability, clinical deployment and potential benefits to working practices and patients.

First, subjects filled an informed consent form. Then they were asked to complete a pre-experimental questionnaire that consisted of basic demographics, familiarity with



**Fig. 4.4** Data entry instances presented on the laptop with (for training) and without (for deployment) grid overlay



**Fig. 4.5** Experimental protocol comprises a pre-experiment questionnaire (left), an introduction video (right, top), the data entry tasks (right, middle), and a post-experiment questionnaire. All experiments are video-taped with the experiment leader taking notes (right, middle). A mobile phone mirrors the Glass screen to check the progress of the task; a remote control is used to switch to the next task when a task is completed. A phonetic alphabet and a summary of Glass commands are provided to the user (left)

cystoscopy and digital mobile devices and opinions regarding the existing (paper-based and image or video capture) methods of data capture. Afterwards, participants watched a short tutorial video developed by us demonstrating the basic operation of Glass, how to navigate the application and how to use the three input modalities.

Participants were then given Glass to wear and asked to adjust it to their comfort before being shown a series of 20 data entry instances for each modality (hereafter called training phase—this was so named as we hoped to specifically gain insights into the learning curve for each modality). Each data entry instance consists of an image shown on a laptop at around 1 m from the participants that corresponds to the view that the Glass application presents via HUD (i.e. bladder schema with the grid overlay), with the addition of one cell that is highlighted and showing a number (Fig. 4.4, left). Participants were then asked to navigate (using a specified data entry modality) via the HUD to the corresponding grid coordinate shown to them on the laptop screen and input data that matched the number shown, with all data being recorded by Glass.

Subjects were then asked to complete a further five data entry instances using each modality in turn in what we named the ‘Deployment Phase.’ This was organised in the same way as ‘Training Phase’ in that each participant was shown a set of images on the laptop and was asked to navigate to the correct cell in the Glass application

and populate that cell with the correct number. In ‘Deployment Phase’ though, the image presented on the laptop did not contain a grid for reference, to more accurately represent the way in which this application would be used in clinical environments (this represented a first step to testing the application in clinical situations but without involving patients; it was considered during planning the protocol that it would not be ethical to involve patients without an initial viability study). The image presented to participants was a copy of the image they could see in Glass but with no grid overlay included; instead, we overlaid a texture taken from a real cystoscope image that would make it harder for participants to relate to that which was visible on Glass (see image#). We felt that this would give us a more realistic idea of which data entry modality performed better in terms of speed to enter data and accuracy as the cystoscopes we observed in use did not include a grid and the limited view offered gave little in the way of anatomical clues with which to orient the user. While the images presented do not entirely represent those seen through a cystoscope, we wished to compare the speed and accuracy between ‘Training’ and ‘Deployment’ phases, and so there was a requirement to keep the size and shape of the images in each phase identical while still increasing the difficulty to the user.

Finally, participants were asked to complete a post-experimental questionnaire. It aimed at eliciting opinions on how fast volunteers were able to learn how to use both Glass and the application and at comparing all three modalities of data entry. Questions covered, for all three input modalities, perception of speed, accuracy, ease of use, acceptability in the clinical environment to clinicians and patients and for fatigue/discomfort experienced. Questions were then asked to gauge the opinion of the extent such an application could improve current methods of data capture and recording and how it could be integrated into current guidelines. Verbal feedback was also elicited. The entire experimental protocol was video-taped for further analyses. The experiment leader was both experienced on the technological side and on the clinical side, and in particular, had experience with cystoscopy which allowed framing of the discussion during debriefing. The entire experimental protocol lasted about 90 min. Details about the cohort are provided in Table 4.1. Except one, none of

**Table 4.1** Cohort comprised eight subjects of various expertise level performing cystoscopy procedures in a UK hospital

| User | Urology practice | Vision aids | Mobile technology use | Use of wearables |
|------|------------------|-------------|-----------------------|------------------|
| 0    | >10 years        | Lenses      | +                     | Fitness          |
| 1    | <5 years         | Contacts    | ++                    | None             |
| 2    | 5–10 years       | None        | ++                    | None             |
| 3    | >10 years        | Lenses      | ++                    | None             |
| 4    | <2 years         | None        | ++                    | None             |
| 5    | <5 years         | Glasses     | +                     | None             |
| 6    | <2 years         | None        | ++                    | None             |
| 7    | 5–10 years       | None        | ++                    | None             |



the participants was involved in the design input. Glass was connected to the hospital Wi-Fi network for speech recognition.

## 4.6 Results and Analysis

### 4.6.1 *Speed, Learning Effect and Accuracy of Entry*

Figure 4.6 reports the time and number of data entry attempts taken to complete the data entry instances in training and deployment phases, for each input modality. The time is given in seconds for the speech modality and in seconds divided by the distance from the centre to target cell in the head movement modalities to normalise for the travel distance. Some users appear as outliers, such as subject 1. In this case, this participant provided continuous verbal feedback during the actual task, even though all participants were instructed to keep feedback for the end (Fig. 4.7).

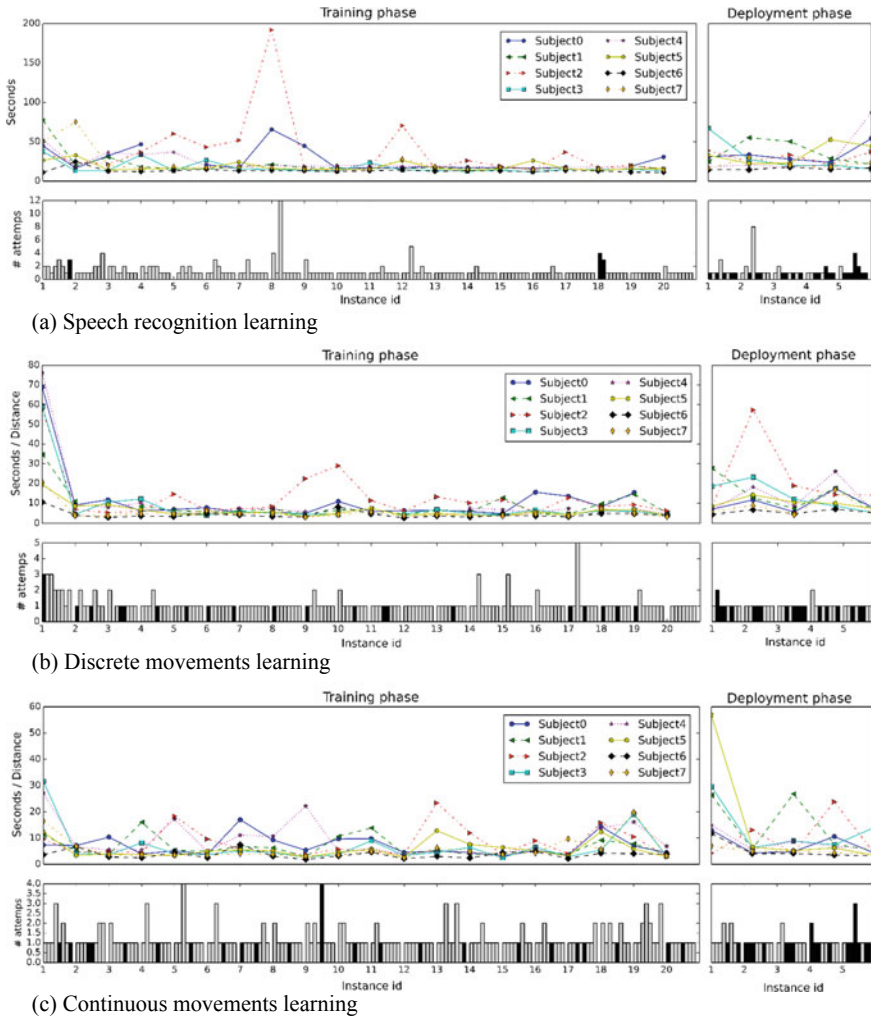
The number of data entry attempts was generally higher at the start of the training phase compared to the end, reflecting increased proficiency at data entry, with the exception of the ‘continuous’ modality, where a ‘lock on’ issue (discussed in ‘Issues Encountered’) explains why the number of attempts does not decrease over time.

Participants became significantly faster over time for the speech and discrete movement modalities, but not for the continuous movement modality (Fig. 4.8). Compounding this, continuous movements were about twice as fast as discrete movements at the start of the training phase, while at the end of the training phase the speed of discrete was slightly faster (Fig. 4.8). Hence, discrete movements required training to increase proficiency, while continuous movements allowed close to optimal speed from the start.

When removing the reference grid in the deployment phase (to better reflect the real condition where the cystoscope does not present such a grid), the time to enter data remained identical for discrete and continuous movements, but increased by almost 50% for speech recognition (Fig. 4.9). This is likely explained by the fact that the cell within Glass can be placed using a visual intuition as to what is closest to the cell highlighted on the laptop. However, when using speech recognition participants had to first visually map where the cell shown on the laptop would be within the grid shown in glass, and only then could say its coordinates.

The accuracy of data entry was very high in the training phase, as participants could undo a cell entry and retry until they were satisfied (which is reflected by more time/attempts to complete the task); however, accuracy decreased in the deployment phase (Fig. 4.10). The accuracy in the deployment phase compares what is the true coordinate of the highlighted cell in the grid-less laptop display with the actual cell participants entered in Glass. Even though the pictures shown on Glass and on the laptop looked different (i.e. more representative of a real cystoscopic procedure), we found out that participants entered a lesion at most two cells away from the true location, with the majority of lesions entered exactly one cell away from the true





**Fig. 4.6** Time (upper plot) and number of attempts per user (lower bars) when using the speech input modality (a), discrete head movements (b) and continuous head movements (c). The groups of eight bars per instance indicate how many attempts each of the eight users made at solving the data entry task. If the bar is white, it means that the task was completed successfully, possibly after several attempts. If the bar is black, the user ultimately failed to enter the data correctly. On the left, the data is shown for the 20 training instances and on the right for the five instances in the deployment phase

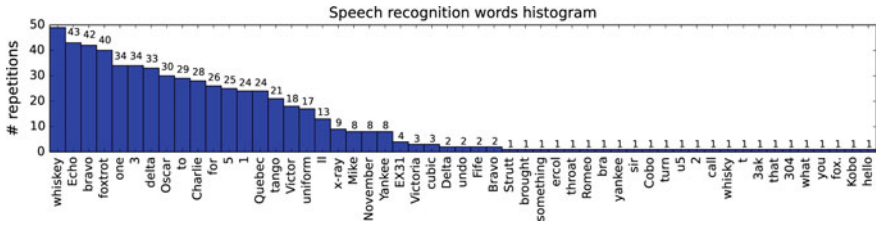


Fig. 4.7 Histogram with the number of identifications per word using the speech recognition modality

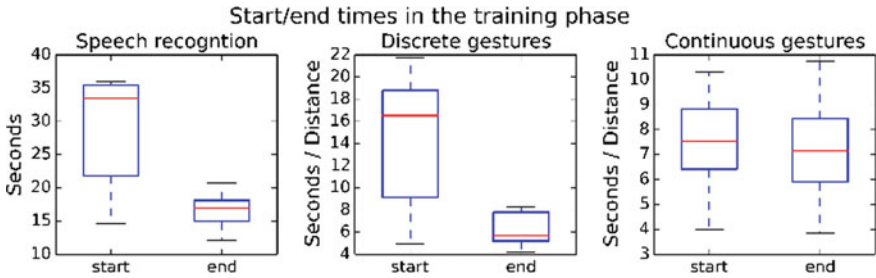


Fig. 4.8 Average data entry time per modality for the first and last 5 instances during the training phase

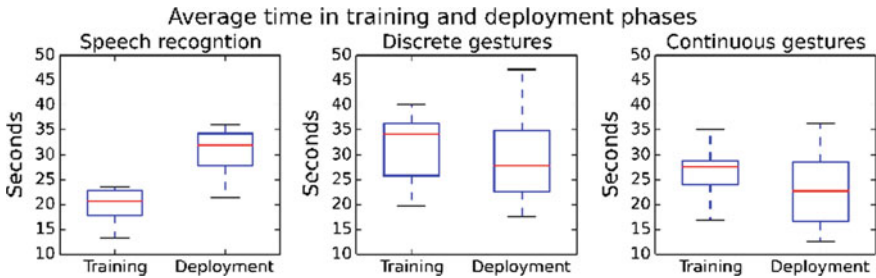


Fig. 4.9 Average data entry time per modality for all the instances during training and deployment phases

location (Fig. 4.11). This indicates a very low inter-user variability in the pinpointing of lesion location, which is important for reproducible data entry.

Speech recognition posed issues as single word recognition was not sufficiently accurate. Figure 4.7 shows a histogram of recognised word occurrence. Ideally, all the words should belong to the phonetic alphabet and numbers 1–5, but the long tail indicates mis-recognised words as well as user errors (e.g. ‘hello’, where the user likely was confused about the data entry procedure). As some errors were frequent (e.g. ‘fife’ instead of ‘five’), the application that was deployed included a substitution table to correct common recognition errors (e.g. ‘II’ instead of ‘2’, ‘for’ instead of ‘4’).

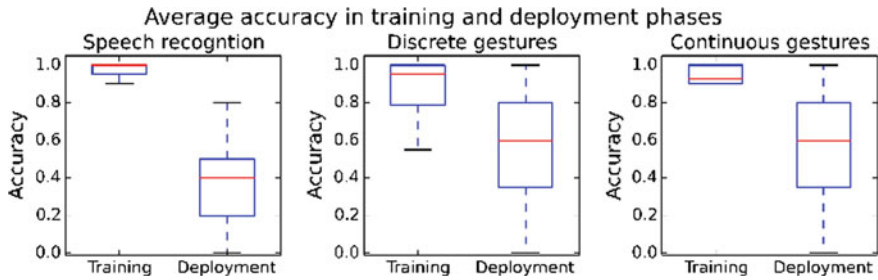


Fig. 4.10 Average accuracy per modality for all the instances during training and deployment phases

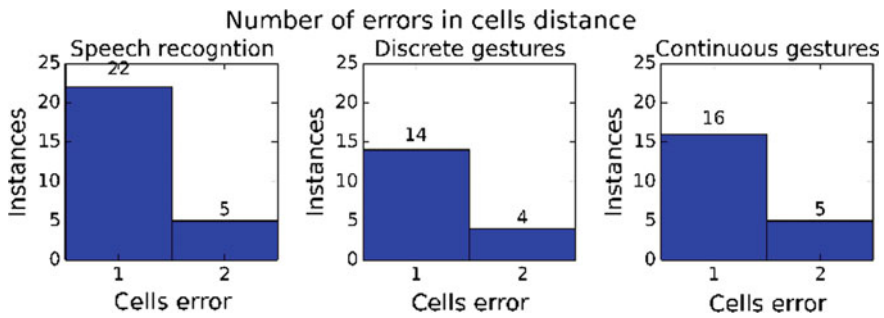


Fig. 4.11 Error in terms of cells distance during deployment phase

### 4.6.2 Current System of Annotation (Paper Based)

Attitudes towards the accuracy and consistency of the current paper-based system of using multiple drawings to identify, describe and track lesions within patient notes were elicited. The highest scores on a five-point scale of poor to excellent were merely satisfactory, very much reflecting the issues with current methodology identified in the genesis of this project. 6 out of 8 participants rated the drawings as either Poor or unsatisfactory on a five-point scale from poor to excellent when asked about the accuracy and consistency of tracking lesions over time by the current paper-based method, with two rating the current system as poor or unsatisfactory on all measures (accuracy and consistency of identifying lesions, tracking lesions over time and locating lesions). 3 out of 5 of the low scores in this regard came from more junior clinicians and two of the more senior clinicians rated the current methods as poor or unsatisfactory across all measures.

### 4.6.3 Preferred Glass Entry Modality

As part of our post-experiment questionnaire, we asked participants to rank the three data entry modalities (Discrete, Continuous and Speech) and give reasons for their choices. We were then able to compare these choices with not only the answers to more specific questions regarding the use of the application but also the raw data collected by Glass regarding learning curves, speed and accuracy related to each modality.

We found that Speech entry was ranked first by four participants (users 0, 2, 3, 4) and Continuous entry by three participants (users 5, 6, 7). Users 0, 2, 3 mentioned in their comments that speed and accuracy were the main motivation behind choosing Speech input with user 3 also mentioning ease of use. Users 4 and 2 commented that Speech input only beat Continuous due to the 'lock on' issue (see below).

Continuous data entry was also chosen first for speed by users 6 and 7. All users that chose Continuous as their first choice mentioned either ease of use or the intuitive quality of this modality.

Discrete data entry was ranked first by user 1. In the comments, provided the 'lock on' issue with Continuous mode was mentioned, as was the fact that Speech relied on proscriptive language and knowledge of the phonetic alphabet. It was also mentioned that the speech recognition was imperfect and could react to other people talking in a clinical environment.

Qualitative data recorded by Glass during testing puts some of these choices in a new light. While Speech was ranked first due to speed and accuracy by users 0, 2 and 3, Speech input was the slowest modality overall.

Even when considering a reduction in entry times of nearly 50% for Speech, it remained at least twice as slow in terms of data entry when compared with Continuous (fastest at the beginning of the experiment) and Discrete (fastest at the end of the experiment).

Continuous was ranked first by three participants, with speed but also ease of use and intuitive controls being mentioned as rationale. This is very much borne out by the data recorded by Glass; Continuous entry was not only the fastest modality at the beginning of the experiment (<50% of the time taken to enter data using Discrete and nearly 20% of the time taken to enter data using Speech) but was also the only modality without a clear learning effect. Speed to enter data remained almost constant throughout for Continuous mode suggesting a data entry style that is easy to understand and quick to grasp.

Discrete data entry was the least favoured, with comments such as 'Hard to use' (User 7), 'Clumsy and slow' (user 6) and 'slowest option' (user 0) highlighting the general user perception of this modality. When compared with the quantitative data from Glass; however, discrete entry can be seen to be twice as fast as Speech throughout the experiment and with a training effect (71% reduction in time to enter data) that placed it as the fastest modality by the end of the experiment.

Speech input was also rated on user perception of accuracy and during the ‘Training Phase’ this was certainly the case, although the difference between the modalities in this respect was small.

When compared to the ‘Deployment Phase’ though, both Discrete and Continuous modalities suffered far less from the lack of grid on the image presented to users via the laptop. While Discrete and Continuous modalities both saw a reduction in accuracy of ~40%, Speech input experienced a reduction in accuracy of >60% making it by far the least accurate modality when being used in a format more representative of that which would be seen in clinical practice. We suggest that this discrepancy in accuracy between Continuous/Discrete on the one hand and Speech on the other may be due to increased visual-spatial cues when using movement to relate to the image. This would account for such a large decrease in accuracy for Speech modality, where having to read coordinates on Glass but not having them visible on the presented image increased the user’s cognitive load which decreased accuracy and increased the time to enter data.

What we can observe here is that user perception does not always match the data recorded by Glass. While Speech input was ranked highly by the majority of the participants in terms of speed and accuracy, it produced results that are objectively slower and (when used in ‘Deployment’ phase) far less accurate. Discrete input was ranked lowest by 5 of the users and only first by 1 (user 1) despite being the fastest mode by the end of training and also the most accurate during ‘Deployment Phase’ (although the difference in accuracy between continuous and discrete was not significant). Despite this, physical discomfort experienced by all but one user while using discrete entry must be accounted for, especially if a device such as this were to be used for prolonged periods.

The difference in accuracy between ‘discrete’ and ‘continuous’ may be explained by the ‘Lock on’ issue whereby users described problems with the selected cell moving while using voice commands to enter the correct icon (see below). It is possible that addressing this issue could see a further decrease in data entry times for continuous mode which, combined with the consistency of speed to enter data and intuitive use, provides valuable information for future application design of this type.

#### ***4.6.4 Potential to Improve Data Recording***

As part of the post-experimental questionnaire, we asked a number of questions related to users’ perceptions regarding the potential of our application to improve on specific areas identified during the design input stage and in our pre-experimental questionnaire. Improvements in the way in which data is recorded during cystoscopy, either during diagnosis or follow-up, appeared to us to be one of the central issues identified in this way.

Participants were asked to use a five-point scale (from ‘Not at all’ to ‘Considerably’) to rate several aspects of the application’s potential. When asked to describe

*'potential to improve the accuracy of describing the site of bladder lesions'* 6 out of 8 of the users selected 4/5 and 1 user selected 5/5. *'Potential to improve the accuracy of tracking progression of bladder lesions,'* saw a similar positive response with 5 out of 8 users selecting 4/5 and 2 users selecting 5/5 (Considerably). This shows a positive response among participants in this regard and reflects a high satisfaction with a digital method of data collection that could be used hands-free while still operating a cystoscope. This is reinforced somewhat by the answers to the question *'To what extent would you find a HUD such as that presented today a detriment to your fine motor control while performing cystoscopy'*. To this question, 4 out of 8 participants selected 1/5 (Not at all), 1 participant selected 2/5, 2 participants selected 3/5 and 1 participant selected 4/5. While only three users suggested that there could be some degree of impact on hand–eye coordination or fine motor control while using the application, this would need to be investigated carefully in future were similar devices to be used in clinical environments to rule out any negative impact on patients.

Inter-user variability was also identified as a drawback to the current paper-based notation (whereby multiple clinicians perform subsequent cystoscopies using individualised annotation styles) and was a key target for improvement when designing the application. When asked, *'After using the application do you feel it has the potential to reduce inter-user variability between personnel performing cystoscopy'* 4 out of the 8 participants selected 4/5 and 3 participants selected 5/5 (Considerably) suggesting that this important drawback in the current method could be addressed successfully by an application of this type. Similarly, when asked about *'potential to improve consistency in the reporting of bladder lesions'* 4 out of 8 participants selected 4/5 and 2 out of 8 selected 5/5 (Considerably).

When asked about *'the potential to speed up the process of note taking after cystoscopy'* the response was more ambiguous; participants 0 and 2 scored 2/5, participants 4 and 7 scored 3/5, participants 1, 3 and 5 scored 4/5 and participant 7 scored 5/5.

#### **4.6.5 Acceptability in Clinical Environments**

The questionnaire assessed acceptability to practitioners and to patients. Acceptability of a HMD used in this manner was high. When asked *'how easy would you find the integration of Glass into your own practice'* 5 out of 8 participants scored 4 or 5 on a 5 point scale from very difficult to very easy. Only 2 out of 8 users felt that it could present issues; importantly physical issues with using the Glass itself were mentioned by both. When asked *'How acceptable would you find the use of a device such as Glass during cystoscopy'*, 6 out of 8 participants scored 4 or 5 on a five-point scale from *'entirely unacceptable'* to *'entirely acceptable'*. Interestingly, none of the users predicted that Glass would interfere with hand–eye coordination or their use of other equipment while working and all felt that Glass would have little or no impact on infection control.

“Acceptability to patients” was estimated by the subjects to be high with 6 out of 8 subjects scoring 4 or 5 on the same five-point scale, although there were concerns raised by three participants regarding the potential impact on doctor–patient interaction of introducing a piece of equipment that ‘sits between me and the patient’ and could present distractions by ‘spending more time looking at the glass than the patient’. Balancing this were the views of 2 participants who felt that it could have an entirely positive impact on interaction, scoring 5/5 in this regard.

#### ***4.6.6 Issues Encountered***

The most common problem encountered was the inability to ‘Lock on’ to cells while using ‘continuous’ modality (all eight participants mentioned this). The current Glass GDK does not notify applications when ‘ok glass’ is detected. It only sends an event when a menu command is detected after that. Consequently, the application is unaware of when the user decided to ‘lock on’ to a cell by saying ‘ok glass’; it only knows when ‘new cell’ is spoken. In the interval between ‘ok glass’ and ‘new cell’, the head-tracking algorithm keeps running, which requires users to keep the head still until they pronounce ‘new cell’. This often meant that the cursor would move from the selected cell, while voice commands were being given, resulting in errors and accounting for the majority of repeated attempts. Despite this, ‘continuous’ was very popular as a data entry method and this represents a major factor when considering the design of such software. Comfort and perception of usability of ‘continuous’ were high, being the second most preferred modality (see above) and this should be weighed against actual efficiency. It would be preferable however to address this issue in future. Another major issue encountered was that of the speed of voice recognition. Five subjects became frustrated at delays between enunciation of a command and its recognition by Glass and comments such as ‘This is not going to be fast enough for Doctors’ were offered by three subjects. Indeed, the same five users raised concerns that Glass itself was not yet responsive enough (in terms of transitioning between screens and responding to voice commands) to introduce into clinical environments.

Three participants suggested that more than one 2D image could be utilised to increase the accuracy of sitting lesions, therefore helping better track lesions over time and as an aid to digital still images recorded by the cystoscope to provide surgeons with more information when planning procedures. Four participants felt that an HMD could also be useful for other tasks, especially in recording notes and accessing patient data/images during procedures and that an application such as this could work very well in conjunction with images recorded by the endoscope by giving more reference points for interpreting the images. It was also mentioned by one of the volunteers that the ability to ‘Map or merge the Glass image onto the cystoscope image’ would be beneficial.

### **4.6.7 Glass**

When asked about Glass itself, there were issues identified with wearing prescription lenses; Glass in the form that we tested does not have a facility for combining with prescription lenses, and so it was sometimes difficult to prevent Glass from moving from an ideal position when resting on top of prescription frames. Those participants using contact lenses reported no issues. One participant had a visual field defect in the eye over which the Glass optic was positioned and this caused a substantial amount of discomfort and difficulty in data entry. This user was concerned about the inability to adjust Glass to use with his other eye and provided the lowest scores across all questions focusing on the system's ability to improve current note-making methods. This user also reported migraine symptoms the following day which they associated with visual issues and neck discomfort from the testing; these are extremely important factors to consider in planning future introduction of such technology into work environments where long usage times and imperfect ergonomics would further exacerbate any physical effects on users. Transient and minor neck discomfort was associated with both 'continuous' and 'discrete' input by five users, citing unnatural neck positions or keeping neck extremely still to avoid moving the cursor inadvertently. One other participant also raised physical issues with Glass, being particularly small in stature they were unable to position the Glass securely on their head and found that it moved out of place during testing. Interestingly, this volunteer provided the second lowest scores in terms of accuracy and potential to improve current note taking suggesting that physical difficulties in using Glass may have a substantial and negative impact on the perception of function and user acceptance. One other issue identified was that of the environment in which Glass is to be used, with three participants mentioning that a blank white wall as a backdrop to patient encounters would be the only way they could use Glass effectively. They indicated that any variation in the backdrop caused difficulty in visualising details and between normal vision and that which appeared in the Glass optic.

### **4.6.8 General Discussion**

There was an overall positive reaction towards using an HMD such as Glass within daily clinical activities and its potential impact on current note taking within the NHS (UK's national health service) Trust it was trialled in. While some of the issues that we aimed to address with Glass such as standardisation of annotation of cystoscopy findings are also targeted by more modern endoscopic technology, there remains a consensus of a continued need for annotated diagrams to be used as an adjunct to HD endoscopic images in the case of cystoscopy; the spherical nature of the bladder in particular often causes difficulty in locating lesions due to lack of reference points. The ability to make notes of this sort during the procedure rather than as an afterthought was considered a positive addition. Cost is another area where HMDs



such as Glass could be beneficial when compared with updating whole endoscopy stacks (estimated at £50,000–100,000 per device). Glass could offer an affordable interim measure to improve notation of findings while also potentially being compatible with future hardware updates due to the universality of its software.

Glass is also suitable for a developing approach to working known as ‘bring your own device’ (BYOD) whereby employees conduct official work on privately owned technology. This approach allows for very flexible working patterns and is already common among clinicians as regards other medical equipment and computing devices, with stringent guidelines for data protection already in place.

Limitations to this study include its size, with only eight subjects, it makes it difficult to estimate general acceptability across the NHS although the generally positive comments are encouraging and bear more investigation. While the study size was small, it did include surgeons and junior doctors across a range of grades, with seven out of eight subjects having experience of other NHS Trusts and four subjects having experience of working in private medicine. The positive reactions from those practitioners having experience with other, more modern, equipment suggest that this system could be suitably integrated to other environments. The time that each participant was able to use Glass was also constrained as we reduced the extent of the testing (to under 90 min) originally planned in order to accommodate the busy working days of the subjects. It was also not possible to trial Glass with patients yet, as this work is an early acceptability study.

A drawback to speech recognition technology as a whole was encountered whereby any word within Google’s dictionary is recognised and the API does not allow limiting words to a specific dictionary (e.g. the phonetic alphabet). Any further development of this application for more extensive notation/annotation would necessitate extensive research into commonly used words or phrases for describing findings. Ideally, it would be possible to make available standardised statements to apply to the image or to refine the speech dictionary.

Other input modalities may become available in future eyewear computers, such as eye tracking. They have not been considered in this project as one particular challenge in cystoscopy is the spherical nature of the bladder. It is not possible to capture the entirety of the bladder in a single image of an endoscope, and therefore, it is not possible to rely solely on gaze direction to point to locations of lesions, as the knowledge of the orientation of the cystoscope within the bladder is missing. Future cystoscopes may be equipped with inertial tracking systems, e.g. to monitor surgeon’s skills [49] which could help in this registration problem.

#### ***4.6.9 Lessons for a Smart Home Environments***

The clinical context has often much more stringent constraints than home assistance, such as rapidity of use, precision, suitability for the sterile environment, and others. In this work, some of the issues we identified with Glass need to be addressed to make the device more suitable for home assistance.

Precision of data entry is important to users. Incorrect recognition of keywords can rapidly become a burden. Our evaluation lasted a short amount of time compared to a daylong use at home, which would only amplify these effects. Therefore, careful optimisation of keywords must be conducted to minimise wrong recognition. Similarly, speed of entry is important, and the right trade-offs must be identified in user studies to maximise acceptance.

Comfort is an important factor which was evidenced with the head gestures which were often noted to be cumbersome. Head gesture entry may not be suitable in the way we proposed in this work. The combination of head motion with gaze tracking may be one way to address this.

Subjective and objective factors must be considered together. In our work, the speech input modality was the preferred one, but was slower and less accurate than the others. Relying on subjective self-reports only would not have uncovered this issue.

Overall, devices such as Glass offer a sophisticated platform to deliver smart assistance. One application domain where Glass may be particularly interesting is in the context of assistance for Parkinson's disease [50].

## 4.7 Conclusion

This chapter examined the potential of Glass to provide an affordable, hands-free data entry device for use by urologists. To our knowledge, there have been no previous studies undertaken investigating the use of HMDs for this purpose. It contributes by showing one approach to systematic lesion annotation, supported by a reference grid, and it provides valuable information on how head gesture sensing and speech recognition is best used in this environment while also presenting feedback from clinicians working with these procedures on a daily basis. The feedback from participants was positive suggesting a high acceptability of HMDs in the clinical environment, with negative feedback focusing on physical issues such as visual defects and the form factor of Glass itself. While the potential for HMDs to improve current note taking systems in terms of consistency and in accuracy of locating and of tracking lesions over time were both viewed positively, as was the potential to reduce paperwork for clinicians, it was felt that the technology was not yet responsive enough in its current form. The approach of providing an image with a grid on which to annotate findings was viewed as a positive contribution. Opinions on acceptability to patients were more mixed, with some concerns raised over the impact on communication with patients but also the potential to improve communication of diagnoses. While it was felt that Glass would have little or no benefit in terms of increasing the speed of diagnosis, it could make surveillance of cancer patients more accurate and could also aid in providing more accurate data to surgeons. Glass was considered by all users to be easily adaptable to infection control guidelines.

'Speech' and 'continuous' were the preferred data entry modalities, rated mainly on ease of use and speed. This was not represented in the data recorded from Glass

which showed ‘continuous’ and ‘discrete’ as being the fastest and ‘speech’ least prone to error of the modalities. ‘Discrete,’ however, was disliked by the majority for being uncomfortable and difficult to use. Comfort and intuitive learning appear to have a very strong influence on perception of speed and accuracy of function.

Future studies could test the system’s impact on ability to use precision equipment and its acceptability in a ‘live’ clinical environment. Side-by-side comparison with handwritten drawings would be required to firmly establish the positive impact of a persistent single annotatable image. Further refinement of the application could include audio notes added to images, compatibility with EHR and medical terminology for speech recognition. Finally, other HMDs could be compared to establish ideal ergonomics.

The source code of the Glass application is available at <https://github.com/sussexwearlab/GlassMedicalDataEntry>.

**Acknowledgements** This work was partly funded by the Austrian FFG project #5766494 “MinI-Attention: Attention Management in Minimal Invasive Surgery”.

## References

1. Cancer Research UK (n.d.) Bladder cancer statistics. Retrieved 03 Mar 2016 from <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bladder-cancer>
2. Lee CSD, Yoon CY, Witjes JA (2008) The past, present and future of cystoscopy: the fusion of cystoscopy and novel imaging technology. *BJU Int* 102(9b):1228–1233
3. Ploeg M, Aben KKH, Kiemeny LA (2009) The present and future burden of urinary bladder cancer in the world. *World J Urol* 27(3):289–293
4. Macmillan Cancer Support (2011) Symptoms of bladder cancer. Retrieved 03 Mar 2016 <http://www.macmillan.org.uk/Cancerinformation/Cancertypes/Bladder/Symptomsdiagnosis/Symptoms.aspx>
5. Zheng C, Lv Y, Zhong Q, Wang R, Jiang Q (2012) Narrow band imaging diagnosis of bladder cancer: systematic review and meta-analysis. *BJU Int* 110:E680–E687
6. James AC, Gore JL (2013) The costs of non-muscle invasive bladder cancer. *Urol Clin North Am* 40(2):261–269
7. Sievert KD, Amend B, Nagele U et al (2009) Economic aspects of bladder cancer: what are the benefits and costs? *World J Urol* 27(3):295–300
8. Yeung C, Dinh T, Lee J (2014) The health economics of bladder cancer: an updated review of the published literature. *PharmacoEconomics* 32(11):1093–1104
9. NICE Guidelines (2015) Bladder cancer: diagnosis and management [NG2]. Retrieved 03 Mar 2016 from <https://www.nice.org.uk/guidance/ng2/chapter/1-Recommendations#diagnosing-and-staging-bladder-cancer-2>
10. RCS (2014) Good surgical practice. Retrieved 03 Mar 2016 from <https://www.rcseng.ac.uk/surgeons/surgical-standards/professionalism-surgery/gsp/documents/good-surgical-practice-pdf>
11. Storz. Image 1: SPIESTM. 2013. 96011020 TP2 SPIES 1.2 10/2013/MFL.E
12. Lux M, Riegler M (2013) Annotation of endoscopic videos on mobile devices: a bottom-up approach. In: *Proceedings of MMSys*, pp 141–145
13. UNISOFT Medical Systems (2016) Retrieved 03 Mar 2016 from [http://www.unisoftmedical.co.uk/urological\\_medicine.asp](http://www.unisoftmedical.co.uk/urological_medicine.asp)

14. van Veelen MA, Snijders CJ, van Leeuwen E, Goossen RH, Kazemier G (2003) Improvement of foot pedals used during surgery based on new ergonomic guidelines. *Surg Endosc* 17(7):1086–1091
15. Wille M, Wischniewski S, Scholl PM, Van Laerhoven K (2014) Comparing Google Glass with tablet-PC as guidance system for assembling tasks. In: *Proceedings of BSN workshops*, pp 38–41
16. Shadiev R, Hwang NSW-Y, Chen N-S, Huang Y-M (2014) Review of speech-to-text recognition technology for enhancing learning. *J Educ Technol Soc* 17(4):65–84
17. Singh G, Nelson A, Robucci R, Patel C, Banerjee N (2015) Inviz: low-power personalized gesture recognition using wearable textile capacitive sensor arrays. In: *Proceedings of PerCom*, pp 198–206
18. Silva ES, Rodrigues MAF (2014) A gesture control system for aiding surgical procedures. In: *Proceedings of ISDEA'14*, pp 287–296
19. Brown-Clerk B, Rousek JB, Lowndes BR, Eikhout SM, Balogh BJ, Hallbeck MS (2011) Assessment of electrosurgical hand controls integrated into a laparoscopic grasper. *Minim Invasive Ther Allied Technol* 20(6):321–328
20. Rousek JB, Brown-Clerk B, Lowndes BR, Balogh BJ, Hallbeck MS (2012) Optimizing integration of electrosurgical hand controls within a laparoscopic surgical tool. *Minim Invasive Ther Allied Technol* 21(3):222–233
21. Harvin G (2014) Review of musculoskeletal injuries and prevention in the endoscopy practitioner. *J Clin Gastroenterol* 48(7):590–594
22. Kim J, Huo X, Ghovanloo M (2010) Wireless control of smartphones with tongue motion using tongue drive assistive technology. In: *Proceedings of EMBC*, pp 5250–5253
23. Jing L, Zhou Y, Cheng Z, Huang T (2012) Magic ring: a finger-worn device for multiple appliances control using static finger gestures. *Sensors* 12(5):5775–5790
24. Ruppert GC, Reis L, Amorim P, de Moraes T et al (2012) Touchless gesture user interface for interactive image visualization in urological surgery. *World J Urol* 10(5):687–691
25. Ward D, Blackwell A, MacKay D (2000) Dasher—a data entry interface using continuous gestures and language models. In: *Proceedings of UIST'00*, pp 129–137
26. Silfverberg M, MacKenzie I, Korhonen P (2000). Predicting text entry speed on mobile phones. In: *Proceedings of CHI'00*, pp 9–16
27. Muensterer OJ, Lacher M, Zoeller C, Bronstein M, Kubler J (2014) Google Glass in pediatric surgery: an exploratory study. *Int J Surg* 12(4):281–289
28. Ponce B et al (2014) Emerging technology in surgical education: combining real-time augmented reality and wearable computing devices. *Orthopedics* 37(11):751–757
29. Russell P et al (2014) First “glass” education: telementored cardiac ultrasonography using Google Glass—a pilot study. *Acad Emerg Med* 21(11):1297–1299
30. Knight HM, Gajendragadkar P, Bokhari A (2015) Wearable technology: using Google Glass as a teaching tool. *BMJ Case Rep* 2015:1757–1790
31. Benninger B (2015) Google Glass, ultrasound and palpation: the anatomy teacher of the future? *Clin Anat* 28(2):152–155
32. Davis CR, Rosenfield LK (2015) Looking at plastic surgery through Google Glass: part 1. Systematic review of Google Glass evidence and the first plastic surgical procedures. *Plast Reconstr Surg* 135(3):918–928
33. Schijven MP, Graafland M, Bemelman WA (2015) Google glass in surgery sharpen your vision. *Surg Endosc* 29(S72):0930–2794
34. Chai PR, Babu KM, Boyer EW (2015) The feasibility and acceptability of Google Glass for teletoxicology consults. *J Med Toxicol* 11(3):283–287
35. Vorraber W et al (2014) Medical applications of near-eye display devices: an exploratory study. *Int J Surg* 12(12):1266–1272
36. Krishnamurthy G (2015) Google glass in intervention radiology-potential applications and limitations. *J Vasc Interv Radiol* 26(2):1051–10443
37. Mentis HM, Rahim A, Theodore PR (2015) Referencing ct scans through a headmounted optical display during laparoscopic surgery. *Surg Endosc* 29(S411):0930–2794

38. Aldaz G et al (2015) Hands-free image capture, data tagging and transfer using Google Glass: a pilot study for improved wound care management. *PloS One* 10(4):e0121179, 1932–6203
39. Albrecht U et al (2014) Google Glass for documentation of medical findings: evaluation in forensic medicine. *J Med Internet Res* 16(2):1438–8871
40. Horak K, DeLand SM, Blair DS (2014) The feasibility of mobile computing for on-site inspection. *SAND* 2014:18291
41. Paterson M, Glass MR (2015) The world through Glass: developing novel methods with wearable computing for urban videographic research. *J Geogr Higher Educ* 39(2):275–287
42. Mauerhoefer L, Kawelke P, Poliakov I et al (2014) An exploration of the feasibility of using Google Glass for dietary assessment. Newcastle University, Computing Science, Technical Report Series, No. CS-TR-1419
43. Scholl PM, Wille M, Van Laerhoven K (2015) Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In: *Proceedings of UbiComp*, pp 589–599
44. Scholl P, Schultes T, Van Laerhoven K (2015) RFID-based compound identification in wet laboratories with Google Glass. In: *Proceedings of WOAR*, Article 13, 5 p
45. Wu T, Dameff C, Tully J (2014) Integrating Google Glass into simulation-based training: experiences and future directions. *J Biomed Graph Comput* 4:2
46. Yang T, Choi YM (2015) Study on the design characteristics of head mounted displays (HMD) for use in guided repair and maintenance. In: *Proceedings of VAMR*, pp 535–543
47. Teixeira JM, Ferreira R, Santos M, Teichrieb V (2014) Teleoperation using Google Glass and AR, drone for structural inspection. In: *Proceedings of SVR*, pp 28–36
48. McNaney R et al (2014) Exploring the acceptability of Google Glass as an everyday assistive device for people with Parkinson's. In: *Proceedings of CHI*, pp 2551–2554
49. Khan A et al (2015) Beyond activity recognition: skill assessment from accelerometer data. In: *Proceedings of UbiComp*, pp 1155–1166
50. McNaney R et al (2014) Exploring the acceptability of Google Glass as an everyday assistive device for people with Parkinson's. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp 2551–2554

**Part II**  
**Activity Recognition and Behaviour**  
**Analysis**

# Chapter 5

## Human Activity Identification in Smart Daily Environments



Hossein Malekmohamadi, Nontawat Pattanjak and Roeland Bom

**Abstract** Research in human activity recognition (HAR) benefits many applications such as intelligent surveillance systems to track humans' abnormal activities. It could also be applied to robots to understand human activity, which improves smart home efficiency and usability. This chapter aims to accurately recognize different sports types in the Sports Video in the Wild (SVW) dataset employing transfer learning. The dataset consists of noisy and similar classes shot in daily environments, not in controlled laboratory environments. Heretofore, different methods have been used and developed for this purpose. Transfer learning is the process of using pre-trained neural networks. The experimental results on different splits of the dataset, size, and pre-trained models show that accuracy of 80.7% is achievable. In another experiment, we have used the famous UCF101 dataset which is collected from YouTube and trained a convolutional neural network (CNN) with Batch Normalization (BN). The achieved accuracy for the test dataset is around 91.2%. One application of the proposed system is to integrate it with a smart home platform to identify sports activities of individuals and track their progress.

**Keywords** HAR · CNN · Transfer learning · Batch normalization · Deep learning · VGG16 · UCF101

### 5.1 Introduction

Progress in human activity recognition (HAR) benefits many applications like intelligent surveillance, robotics, autonomous vehicles, smart homes and assisted living. In the context of smart homes/environments, it is very crucial to identify human actions to monitor and keep record of the type of the action in the first place and also to avoid dangerous scenarios such as falling or colliding with objects. An example of the first scenario would be activity-type identification and recording in smart homes, sports halls or gyms. The second intention of HAR can be applied to care homes to help elderly living healthier lives. The focus of this chapter is on the first scenario

---

H. Malekmohamadi (✉) · N. Pattanjak · R. Bom  
Institute of Artificial Intelligence, De Montfort University, Leicester LE1 9BH, UK  
e-mail: [hossein.malekmohamadi@dmu.ac.uk](mailto:hossein.malekmohamadi@dmu.ac.uk)

© Springer Nature Switzerland AG 2020  
F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications  
and Networks, [https://doi.org/10.1007/978-3-030-25590-9\\_5](https://doi.org/10.1007/978-3-030-25590-9_5)

where the proposed system is capable of identifying different activities recorded in uncontrolled environments and can be reproduced by using off-the-shelf hardware and without any specialization.

HAR on its own can be subdivided into: sensor-based and video-based. Sensor-based HAR focuses on analyzing data gathered by sensors such as GPS and Bluetooth as addressed in Chen [9], Lin et al. [23], and Wang et al. [44], while video-based HAR focuses on video or image data as addressed in Hongeng et al. [17] and Ni et al. [27]. Human activities can be grouped into four categories: actions, gestures, group activities, and interactions [2].

Action can be classified as a single person performing one or more gestures in which gestures are movements of the body. Interaction is classified as activities that include multiple persons. Group activities are those activities performed by a group of people. The recognition part is the most complex part of any automated HAR system. One problem for a computer to recognize human activities in videos is caused by variation in the video environment [30]. The different illuminations, occlusion of objects, camera angles, background noise, and video quality are factors that affect the accuracy of the computer recognizing human activities. Among different methods for recognition, deep learning is a very interesting choice for researchers in HAR.

Although the accuracy of recognizing human activity in videos has progressed over the past, mainly affected by improvements gained in the power of parallel computing as well as the high amount of data available, thanks to the Internet, these improvements allowed researchers to develop high computational methods to recognize human activity in videos. Convolutional networks were almost reinvented in Krizhevsky et al. [21] where authors used the improvements in parallel GPUs combined with this old technique to design the famous AlexNet model. This model was a convolutional network that participated in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) and won that competition. This was a huge step compared to previous results.

The dataset used in this chapter is the Sports Videos in the Wild (SVW) dataset [34], which consist of 4206 videos in 30 different classes. The videos are captured using a smartphone app called Coach's Eye. Every class has around 100 videos with lengths from 3 s to 1 min. This dataset is very challenging because the videos differentiate in orientation, lighting, stability, viewpoints, and many other factors. In this chapter, the data will be split into training and testing sets, but will also be tested using an external dataset. The other dataset chosen is the popular UCF101 dataset [38], which is a dataset containing different sports videos divided into 101 sports activities from YouTube. Both datasets have several corresponding sports activities, which allows this chapter to test the model on an independent data source. Other datasets like the Weizmann [7] or KTH [35] have not been selected in this chapter since they are created in controlled environments. These datasets contain activity videos with limited (one) viewpoint hence they are not good candidates to represent the abundance of videos in the internet.

UCF101 is one of the largest datasets in HAR benchmarks. It was introduced by Soomro et al. [38]. UCF101 contains 13,320 video clips which their lengths are between 1.06 and 71.04 s, the resolution is  $320 \times 240$  pixel, and the frame rate is



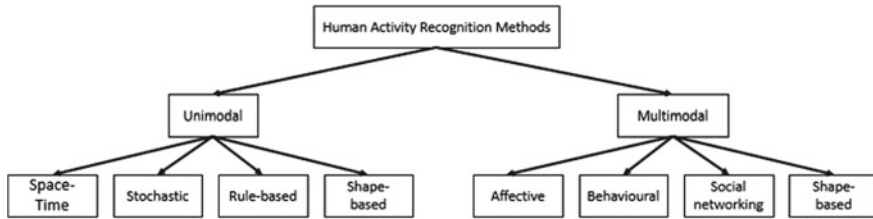
25 fps. It consists of 101 activity classes such as archery, basketball, diving, playing piano, and others. When the dataset was first introduced, the performance baseline was 43.9% by standard bag-of-words method using the implementation by Marszałek et al. [26]. It implies that UCF101 is one of the most challenging datasets of HAR tasks. It consists of a large number of classes, a large number of video clips, and also a large number of different contents of each video clips.

The focus of this chapter will be on improving the accuracy of HAR by implementing deep learning methods applying to the aforementioned datasets. There are many other factors that could have been chosen to focus on, but accuracy is one of the most important factors. Another factor that was considered was training time, but the improvements in GPUs over the years would suggest that this factor is solved over time, which makes it less attractive for us to be studied here. The reason why deep learning has been chosen is also affected by the improvements seen in the field of HAR. The improvements in GPUs offer more possibilities and their easy access to the public makes it an interesting technique to be applied to HAR. Another reason can be found in the fact that deep learning enables machines to work without the involvement of human experts. The training time of a model for HAR can be long hours, but when a model is thoroughly trained, it can be applied relatively quick which makes it ideal to be implemented in automated systems like robots and camera surveillance.

Even though a human can understand objects in an image easily, building a machine to understand image or video such as image classification is challenging. This is because a computer understands image or video as an array of numbers. Changing viewpoints, illumination, deformation, occlusion, and background clutter can change the values in an array of number. Writing a program as a procedure is not a straightforward way to achieve this task. However, convolutional neural network (CNN) model can overcome this problem. CNN is currently one of the most machine learning techniques that play a major role in computer vision. It can convolute a large number of image array to select only useful information (feature extraction) for neural network to classify image features. However, when training CNN model with a large number of dataset, it is challenging to build models to achieve a high rate of accuracy for the first time. It is possible to have overfitting. Adding Batch Normalization [19] is one of the useful techniques that we can improve the model accuracy. In another experiment, UCF101 dataset is used with Batch Normalization (BN). This concept is also covered and applied in this chapter to the famous UCF101 dataset.

## 5.2 Background

There have been many attempts to identify different categories of human activities in videos. The authors of Aggarwal and Cai [1] differentiate unique methods based on the view of the camera: single view or multi-view. In the same year, another researcher divided the different methods based on the video data: 2D or 3D [15] although the above researchers were of course not aware of new techniques which could add



**Fig. 5.1** Different methods used for HAR categorized in unimodal and multimodal methods as designed in Vrigkas et al. [42]

more dimensions to classify the different human activity recognition methods. Over the years, many other different categorization models have been developed such as Jaimes and Sebe [20], Pantic et al. [29], but another method developed by Vrigkas et al. [42] attracted our intention to develop our system presented in this chapter. The reason for this originates from the fact that this model classifies methods into multiple subcategories and includes categories for some of the recent HAR models. This classification model is schematically described in Fig. 5.1.

The first split is the difference between unimodal and multimodal methods. Unimodal data contains one modality. Multimodal methods use data from multiple different sources to combine them in order to improve the success of HAR. This chapter focuses on 2D videos only, which means that all the multimodal methods will not be discussed. The unimodal data can be split up in four different subcategories: space-time, stochastic, rule-based and shape-based. Firstly, the space-time methods are methods that approach human activity recognition problems as a 3D space-time problem instead of multiple 2D space problems. When looking at videos, it is quite easy for a human to recognize a sport from a single image or frame. A computer could do the same, but it would recognize a sport based on a single frame and thus neglects the information stored in consecutive frames over time. Different space-time methods have been developed from focusing on optical flow, in which the optical flow represents the motion of an object in a single frame over time in multiple frames by using a so-called displacement vector that shows the movement of a pixel from the first frame to ones following up [12, 48]. Other space-time methods consisted of methods that used local features to classify video clips [35].

Secondly, stochastic methods focused on using more statistically orientated methods and involved prediction. A popular approach was the use of Hidden Markov Models [6, 49]. Thirdly, the human silhouette is always present in a video with humans. That is one of the reasons for researchers to focus on the silhouette and more specifically on the limbs, to understand different human activities. This can be classified as shape-based methods. These methods have been approached again in different ways from using a method called bag-of-rectangles, which models a human silhouette in rectangular boxes [18]. The most known progress in this method category was done in Wang et al. [45], which introduced the use of depth cameras to the field of human activity recognition, which added a new dimension to the current RGB videos. Nowadays, off-the-shelf devices like Kinect or Intel RealSense are

popular options to capture color plus depth videos. Finally, the last subcategory is the category of rule-based methods. These methods focus on applying rules to describe activities. Activities are split up into smaller subactivities, which later can be used to build up other activities [24]. Most methods applied to recognize human activities can be classified using the four above-described methods although there are lots of interactions between the subcategories, meaning that some methods cannot be purely classified as for instance space-time or shape-based methods.

Most of the aforementioned HAR methods are based on two steps: feature extraction and classification. The first step of feature extraction is probably the most important step in HAR. It consists of extracting features from the frames of the inputted videos. These features represent the essential information in an image to recognize a class (in this case, a sports activity). Facial recognition relies on features that are more distinct or present compared to human activities. In a face, the main features are, for instance, the mouth, nose, and eyes, but for human activity recognition that is completely different. Essential features in human activity recognition are, for example, edges and corners, which are already a lot harder to be recognized and are of a more abstract dimension. In the field of computer vision, there are two main groups of features detection methods: handcrafted methods and the deep learning (feature learning) methods.

The field of handcrafted methods focuses on identifying the features as mentioned above. One of the first methods used histogram of oriented gradients (HOG), which describes shapes by looking at the distribution of gradients in pixels [12]. Other known methods are scale-invariant feature transform methods (SIFT) [25] or local binary pattern methods (LBP) [28]. These handcrafted methods are essentially different from feature learning methods because they need the influence of experts to operate these methods, while feature learning uses raw data and recognizes patterns or features on its own without the influence of experts. The downside of feature learning is that it needs relatively large datasets compared to the handcrafted features [11]. Although, the performance of the handcrafted feature methods is reaching the same (or higher) accuracy levels compared to the feature learned methods, the availability of big data in the digital age and recent advancements in deep learning research are the main motivations for this chapter to focus only on the feature learning methods. If these methods improve, there is a possibility of using this technique in real time, which could be applied to surveillance or other applications (without the need of experts). This means that the handcrafted methods are not within the scope of this chapter, but the focus will be on feature learning methods.

### 5.3 Experimental Results

In this section, we describe our model selection, experiments, and the results for HAR in the SVW dataset followed by another experiment for UCF101 dataset.

### 5.3.1 Experiments on SVW Dataset

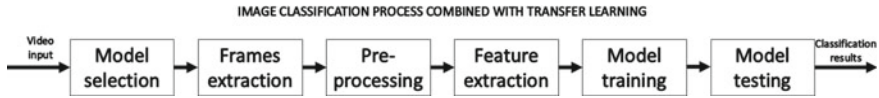
There are four different studies done using different methods, to improve the accuracy of HAR using the SVW dataset. The first attempt was done by the authors/creators of the SVW dataset [34]. All of their methods can be classified as handcrafted feature extraction methods in both the stochastic category and space-time category. Their best attempt used HOG and reached an average accuracy of 61.53%. As stated, earlier handcrafted methods are neglected for this chapter, which makes the other three studies of more interest. The other three studies used deep learning feature extraction methods. One of the study groups looked only at the spatial dimension and neglected the temporal dimension, with the best result containing an average accuracy of 82.51% [31]. They used the first 180 frames of the videos as input and used a convolutional neural network architecture based on the AlexNet module. The authors also wrote another paper in which they used a method to extract the keyframes describing the action based on the highest saliency value, which gave them an improved average accuracy of 84.82% [32]. The final study used different deep learning architectures also including the 3D-CNNs, which are convolutional neural networks that are able to look at the temporal information by using multiple consecutive frames as input. They came up with an average accuracy of 84.30% by using the pre-trained InceptionResnetV2 model [3].

Deep learning demands powerful CPUs and GPUs that lead us to choose Google virtual machines for our first set of experiments. The full specifications of the Google virtual machine can be found in Table 5.1. Expectedly, the main programming language used in this chapter is Python. Some shell scripts have been used to split the data, but most of it is done in Python. The main libraries used in python are OpenCV (extracting and saving frames), Sklearn (model results), Keras (model building), and Matplotlib (plotting results).

The experiment can be interpreted as an image classification process combined with transfer learning elements. Figure 5.2 shows the 6 steps of the image classification process used in this experiment. Prior to any setup, it is key to pick the models first. The reason for this is the fact that each model is different and has specific advantages and disadvantages. All models chosen have been trained on the ImageNet dataset. The ImageNet database is a database containing 14 million images. All the models are trained on 1000 categories of this dataset. This is why for each

**Table 5.1** Hardware elements of the Google virtual machine used in the first set of experiments over SVW dataset combined with a description of those elements

| Hardware | Description             |
|----------|-------------------------|
| CPU      | 2.60 GHz Intel Xeon (R) |
| RAM      | 13 GB                   |
| GPU      | Nvidia Tesla K80        |
| VRAM     | 5 GB                    |
| HDD      | 100 GB                  |
| OS       | Windows server 2016     |



**Fig. 5.2** An overview of each of the six steps in the image classification process combined with transfer learning

of the next model the fully connected layers will be removed and fitted with a new fully connected classifier that classifies into 30 categories instead of 1000 categories. Three models have been selected based on their performance and accuracy scores on the ImageNet dataset as shown in Table 5.2.

One of the models chosen for this chapter is the VGG16 [36] which has 16 layers. This model is relatively small compared to the other models containing hundreds of layers. This model can be seen as an improvement to the AlexNet model. The main reason for choosing this model comes from the fact that it demands a relatively low hardware power. The other selected model is InceptionV3 [40]. It is the third model in a line of series (InceptionV1, InceptionV2). While the VGG16 model and other convolutional neural networks focus on the depth of the model, it was the Inception team that used all kinds of measures in the model to improve accuracy. Nevertheless, the InceptionV3 model was already containing 48 layers. This model showed almost a similar accuracy performance on the ImageNet dataset compared to the Inception-ResNetV2 which is why it was added to the list of models in this chapter. In terms of complexity as shown in Table 5.2, the InceptionResNetV2 can be considered the most complex with its ~56 M parameters. All of these models can be compared to a few other statistics as well. Table 5.2 shows that the InceptionResNetV2 is the best performing model on both top 1 and top 5 accuracies tested with ImageNet dataset. However, the SVW dataset is a complex dataset which could mean that the other models are going to outperform the InceptionResNetV2. When looking at the high number of parameters for each model, it is clear that all models are extremely complex.

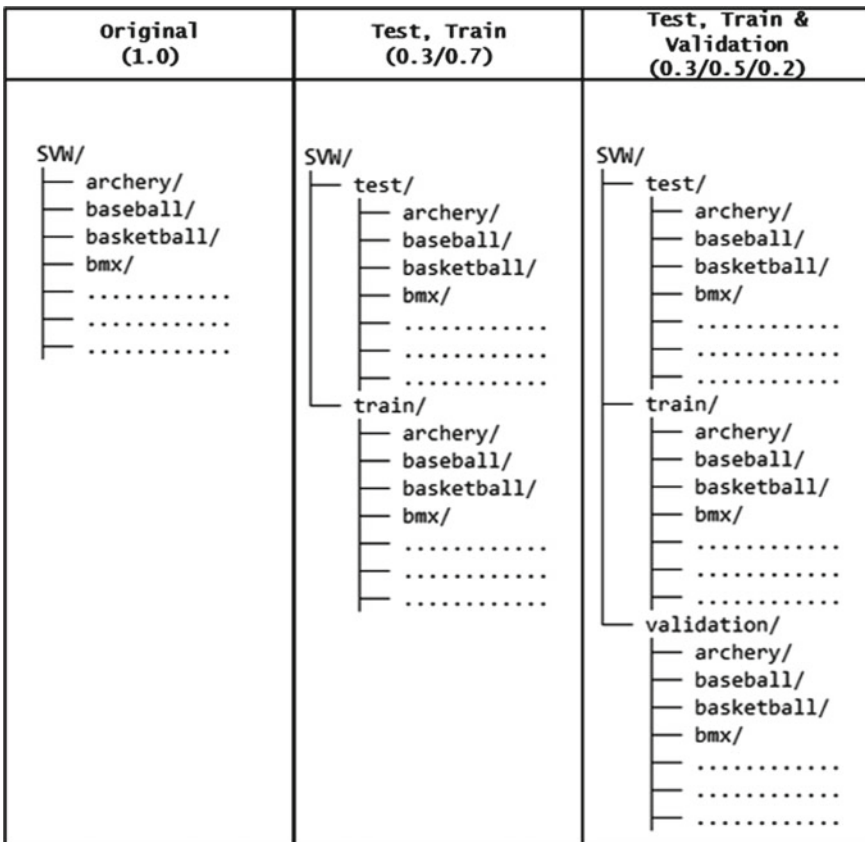
As stated earlier, the SVW dataset consists of the 30 classes/sports activities. The average number of videos for each of the datasets is 128, but there are two classes with a relatively large size. Volleyball and swimming contain both over 240 videos while the next highest class contains 190 videos. This means that these classes contain a lot more information which could influence the final results. The total number of videos is 4206, which is divided into training and testing data. The creator of this dataset made three different splits of 70 and 30% testing data, which should be used

**Table 5.2** Three chosen transfer learning models compared on 3 different statistics, showing their accuracy performances on the ImageNet dataset and number of parameters

| Model name        | Top 1 Acc | Top 5 Acc | Parameters |
|-------------------|-----------|-----------|------------|
| VGG16             | 0.715     | 0.901     | 13,357,544 |
| InceptionV3       | 0.788     | 0.944     | 23,851,784 |
| InceptionResNetV2 | 0.804     | 0.953     | 55,873,736 |

to make the results of this chapter suitable for benchmarking. Although one new split has been added in this chapter, the training data is divided into another dataset: the validation dataset. This dataset will be used to tune some hyperparameters. If this was done on the test dataset, this would have biased the results and would have created a model that would capable of getting a high accuracy on the test dataset, but probably a poor accuracy in general, thus creating a memorized model instead of a generalized model.

These steps create the following folder output (see Fig. 5.3). This figure shows the stages in the frame extraction phase. In the final stage, you can see the training (50%), validation (20%), and test (30%) dataset partitioning. In the above example, each of the class subfolders contains videos. However, the transfer learning process only works with images. All videos have a frame rate around ~30 fps, which would result in a huge dataset if all frames were used. Subsequently, four different sizes of data have been used to sample at 1 frame per 1–4 s. For instance, in one dataset one



**Fig. 5.3** Splitting the data in train (50%), test (30%), and validation (20%) datasets, show in three steps

frame selected every 2 s. This means that a video of a length of 6 s returns 3 frames. This incorporates the fact that the videos with more frames will have more effect on the training process, simply because their frame proportion in the training dataset is higher. This could result in the selection of frames that contain no information about the sports activity. An overview of which model is capable of handling which dataset size is shown in Table 5.3.

Figure 5.4 shows some of the frames of the training dataset of the archery class to show the complexity of the SVW dataset. The displayed frames in Fig. 5.4 contain different shapes, environments, and lighting settings. Each of the models used in this chapter has its own preprocessing steps. Some models work with an input between  $-1$  and  $1$ , while other models work with values with an input between  $0$  and  $255$ . All three models use the same kind of input described in the format of  $(Image_{Width} \times Image_{Height} \times \#Channels)$ . The number of channels is set to  $3$  for each model because every model uses one value for each red, green, and blue. For grayscale images, this is set to  $1$ . The image width and image height of the InceptionV3 and InceptionResNetV2 are set to  $299,299$ . While the VGG16 uses an input of  $224,224$ . These values have not been chosen randomly, but represent the values used by the authors of the models. In this chapter, these values have not been changed. So far, the specific models have been chosen and the data has been preprocessed. The next step is to extract the features.

All of the employed models have been trained on the ImageNet dataset. This means that there is access to the features used to identify different classes in that dataset. Those features form the power of transfer learning. The process of using those features for the SVW dataset can be explained with the VGG16 model. All

**Table 5.3** An overview of which model is capable of handling which dataset size

| Model             | 1 s | 2 s | 3 s | 4 s |
|-------------------|-----|-----|-----|-----|
| VGG16             | Yes | Yes | Yes | Yes |
| InceptionV3       | No  | No  | No  | Yes |
| InceptionResNetV2 | No  | No  | Yes | Yes |



**Fig. 5.4** An overview of the differences between frames in the dataset by looking at the archery sport activity



models consist of a base model and a classifier. At the end of the base model are the last feature maps, also called bottleneck features. The first step done in the transfer learning process is removing the classifier from the whole model which means only keeping the base model. Then, predictions are made using the features from the base model. These predictions are then fed into a new classifier specified for 30 classes; to represent 30 types of HAR in the dataset instead of the 1000 classes in the ImageNet dataset. In this experiment, the classifier used is the standard classifier used for each model as specified in Keras. The classifier is trained using the training dataset (50% of the original dataset) but also the validation dataset (20% of the original dataset), which is used to make some small adjustments in the hyperparameters. There are also some other parameters which have been kept constant or the same for each of the models.

The most important parameters are: (a) Batch sizes of 32,128; (b) Epochs are 200, 1000; (c) Optimizer is Rmsprop and (d) Loss is categorical cross-entropy. Leaving these parameters constant simplifies the problem, but some parameters had to be changed for some models. Interested readers are encouraged to refer to textbooks such as [16] for better understanding of these parameters. The number of epochs has been set to the number 200 for the InceptionV3 and InceptionResNetV2, which results in all models trained until they hardly improve. The number of epochs for the VGG16 is set to 1000 because after 200 epochs the accuracy graph still shows an increasing line. The batch size could have been increased but is limited by the computational limits. This means that this is set to 32 for InceptionResNetV2 and InceptionV3, but altered to 128 for the VGG16. The optimizer is a common optimizer and the loss function chosen is one suited for a multi-class problem and the shape of the data. The model is tested by making predictions on the test data. Then, the accuracy is calculated and the results are displayed in a classification in Fig. 5.5 and confusion matrix in Fig. 5.6. The classification report shows how the model performed by looking at a few different measures like accuracy, recall, precision, and F1-Score. The confusion matrix shows the recall by comparing the true label and the predicted label, but can also be used to identify false positives, which can help in getting a complete understanding of the model performance. The different measures together form a better representation of the classifier performance than accuracy only. The confusion matrix shows some more depth in which classes have been correctly predicted, but also the classes that have been wrongly predicted. In Fig. 5.6, the four highest miss-classified videos are circled in red and are further investigated in an attempt to understand why the model has wrongly classified these frames.

Table 5.4 shows the test accuracy results for each split of each model while changing the input data size. The “X” marks that the specific model was not capable of handling that amount of data as earlier concluded in the frame extraction phase. The first thing that can be concluded is that the second split scores best compared to the other splits. Another thing that can be concluded is that the InceptionResNetV2 scored 80.3% at the highest accuracy, InceptionV3 an accuracy of 79.2, but the VGG16 scored a test accuracy of 80.7%. This 80.7% was reached with 1 frame every second and in the second split. In total have 21 runs been done with different



|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| archery      | 0.83      | 0.89   | 0.86     | 496     |
| baseball     | 0.80      | 0.65   | 0.72     | 1022    |
| basketball   | 0.55      | 0.48   | 0.51     | 398     |
| bmx          | 0.88      | 0.81   | 0.84     | 395     |
| bowling      | 0.85      | 0.98   | 0.91     | 427     |
| boxing       | 0.90      | 0.83   | 0.86     | 998     |
| cheerleading | 0.88      | 0.84   | 0.86     | 793     |
| discusthrow  | 0.51      | 0.67   | 0.58     | 359     |
| diving       | 0.82      | 0.95   | 0.88     | 380     |
| football     | 0.64      | 0.74   | 0.69     | 454     |
| golf         | 0.93      | 0.45   | 0.60     | 511     |
| gymnastics   | 0.72      | 0.76   | 0.74     | 624     |
| hammerthrow  | 0.72      | 0.61   | 0.66     | 370     |
| highjump     | 0.51      | 0.70   | 0.59     | 291     |
| hockey       | 0.92      | 0.83   | 0.87     | 887     |
| hurdling     | 0.56      | 0.50   | 0.53     | 370     |
| javelin      | 0.49      | 0.53   | 0.51     | 318     |
| longjump     | 0.53      | 0.57   | 0.55     | 298     |
| polevault    | 0.68      | 0.65   | 0.67     | 297     |
| rowing       | 0.98      | 0.92   | 0.95     | 662     |
| running      | 0.39      | 0.44   | 0.41     | 549     |
| shotput      | 0.65      | 0.60   | 0.62     | 350     |
| skating      | 0.60      | 0.84   | 0.70     | 306     |
| skiing       | 0.91      | 0.86   | 0.88     | 500     |
| soccer       | 0.46      | 0.64   | 0.54     | 221     |
| swimming     | 0.97      | 0.98   | 0.97     | 3999    |
| tennis       | 0.86      | 0.84   | 0.85     | 572     |
| volleyball   | 0.75      | 0.80   | 0.77     | 1097    |
| weight       | 0.83      | 0.85   | 0.84     | 870     |
| wrestling    | 0.94      | 0.94   | 0.94     | 1816    |
| avg / total  | 0.82      | 0.81   | 0.81     | 20630   |

**Fig. 5.5** Classification report of the best performing model, which is the VGG16 on the second split when taken 1 frame every second



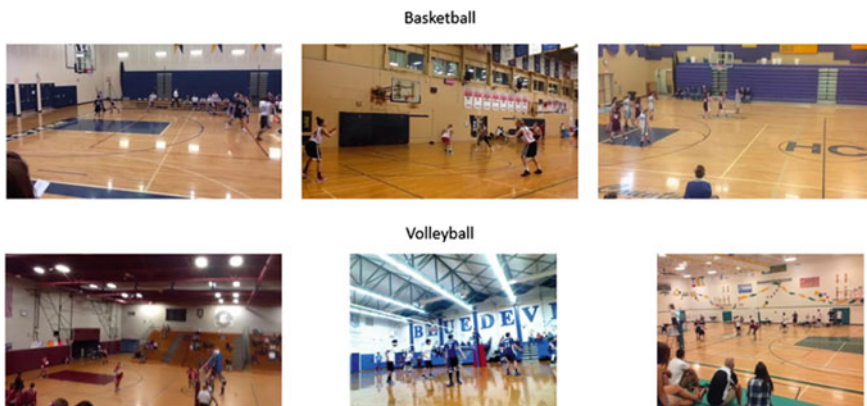




**Fig. 5.7** A few examples of frames containing no sport activity showing: empty grass fields, a blurry frame, and a hand blocking the camera

can be drawn when zooming in at the accuracy and loss graphs, classification report and confusion matrices of the best model.

The first misinterpretation of the model is the fact that basketball frames have been identified as volleyball frames. With a score of 0.42, this is a rather high number, which can be seen as the biggest reason that basketball scores a low score of 0.48 in its own class. You would think that this could be possible because both classes are a team sport played with a round ball that probably stands out in the frames. When looking in the dataset (see Fig. 5.8), the first thing that immediately stands out is that



**Fig. 5.8** Similarities between the basketball and volleyball class, shown by 3 frames of each class

both sports are played in the same environment. Figure 5.8 compares the two with some frames taken from the dataset. The main similarity that can easily be spotted is the environment in which both sports are played. A lot of the frames are taken from videos recorded in indoor halls which produce similar frames. However, there is also another thing that was quite easily spotted when browsing through the frames. This has been shown in Figs. 5.8 and 5.9.

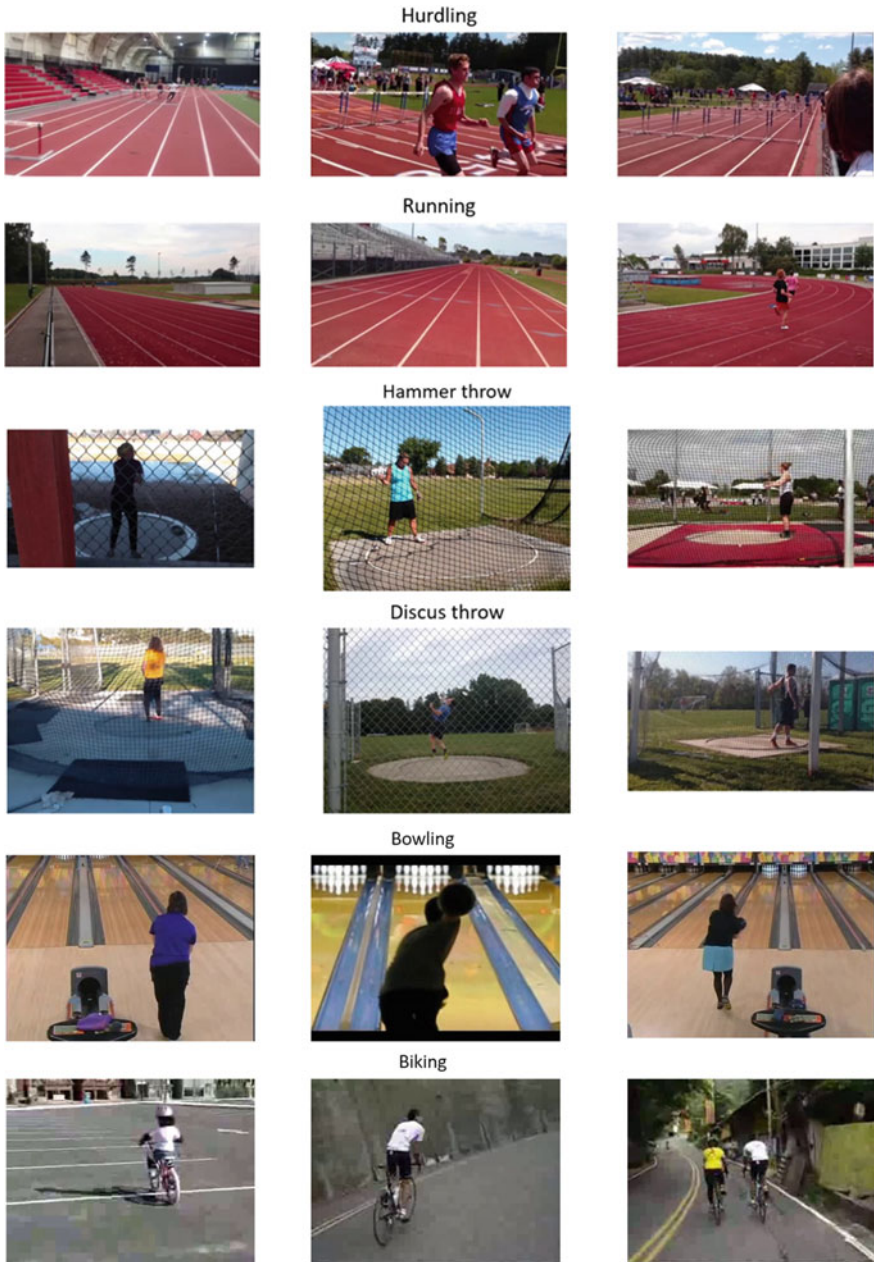
The basketball hoop is definitely a strong presence in some of these frames and could probably easily be a cause of the marking basketball videos as volleyball, because of its strong rectangular shape. Another misclassification occurs between the running and hurdling class. This seems to be caused by the fact that there is a high similarity between the datasets based on where they are recorded. A lot of the running videos contain the same type of cinder track with the same significant colors and white lines (see Fig. 5.10). There is also some misclassification between the hammer throw and discuss throw classes (shown in Fig. 5.10). Both sports are based around throwing an object. However, when closer looking at the frames it is probably not the throwing aspect causing the misclassification, but more likely the presence of netting and the circle they stand on, which is highly visible and thus present in both classes.

As mentioned earlier, the UCF101 dataset is a dataset existing out of 101 sports activities. Thirty of those 101 sports activities have been selected to be used for testing. From the thirty selected, there are about 10 of those classes that have similar videos compared to the SVW dataset. However, the videos are clearly different from the SVW sports dataset which makes it very interesting to test the model on this external dataset. The results of the UCF101 test dataset are poor. When zooming in on the three highest scores, there are some explanations found that could be the cause of the high scores. The first one is the combination of biking and bowling, which



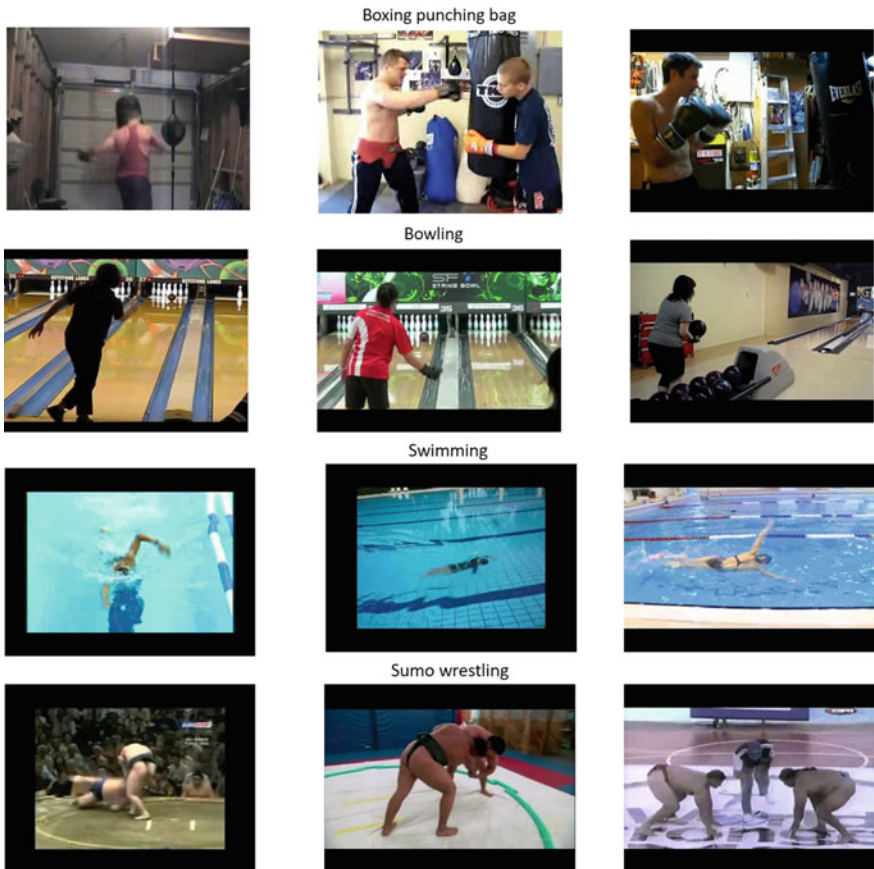
**Fig. 5.9** Basketball hoop (circled in red) in the volleyball class, shown by 3 frames of each class





**Fig. 5.10** Similarities between the hurdling and running class, shown by 3 frames of each class (row 1–2). The similarities between the hammer throw and discus throw class, shown by 3 frames of each class (row 3–4) and an example of the silhouettes of bikes compared to people that playing bowling (row 5–6)

can be seen in Fig. 5.10. The figure shows an example of the silhouettes of bikes compared to people that play bowling. Both seem to have a similar silhouette that causes confusion for the network. Another high score is seen between the boxing and bowling class. Figure 5.11 shows that the silhouette of the boxing class is sometimes similar to the bowling class. The gloves of the boxer also contain similar shapes to that of a bowling player holding a bowling ball. Another score shows similarities between the sumo wrestling and front crawling class. These classes do not show many similarities. The swimming pool has a lot of distinct patterns and blue colors of the water. The sumo wrestling area contains a visible circle in the middle with a lot of people around it. However, there is one thing that does show some similarity, which is the black boxes around the images. From both classes, a lot of images have got these black boxes, which can be seen in Fig. 5.11. Furthermore, image regions filled with skin color can be identifiers for these misclassified activities.



**Fig. 5.11** Silhouette of the boxing class is sometimes similar to the bowling class (first two rows) and similarities between the sumo wrestling and front crawling class (last two rows)

### 5.3.2 Experiments on UCF101 Dataset with BN

The experiments in this section are based on UCF101 dataset where a sample footage of this dataset is presented in Fig. 5.12. Batch Normalization is one of the most well-known techniques to improve the speed of training of deep learning models. It was introduced by Ioffe and Szegedy [19]. When training deep learning models especially using Stochastic Gradient Descent (SGD) as an optimizer, it requires careful consideration of using learning rate and initial values of model parameters. Changing input data distribution, which cannot be avoided in a practical way, in each layer can also possibly cause “internal covariate shift” [19]. Each model layer needs to adapt itself continuously for the change of data distribution. If the distribution changes too much, a deep learning model will experience a difficulty to find a lower loss function value. As a result, the deep learning model will not be able to be trained successfully. Adding BN into deep learning models can prevent a large change in data distribution for each model layer [10]. Therefore, Batch Normalization is suggested implementing to overcome this inevitable problem.

Batch Normalization was found from finding a method to reduce the internal covariate shift. The study was started by fixing the input distribution for each model layer when training is run. Whitening input data [19] can provide a fixed input distribution. It is an original concept that is used to reduce the internal covariate shift. However, whitening input data can be made simplified by normalizing input data. In order to train a deep learning model with an SGD optimizer, mini-batches is also introduced when normalizing input data. The algorithm of Batch Normalization proposed by Sergey Ioffe and Christian Szegedy is presented the Algorithm 1.



Fig. 5.12 Footage of UCF101 benchmark



**Algorithm 1** Batch Normalization [19]

Input: Values of  $x$  over a mini-batch:  $\mathcal{B} = \{x_{1...m}\}$

Parameters to be learned:  $\gamma, \beta$

Output:  $\{y_i = BN_{\gamma,\beta}(x_i)\}$

Mini-batch mean:  $\mu_\beta \leftarrow \frac{1}{m} \sum_{i=1}^m x_i$

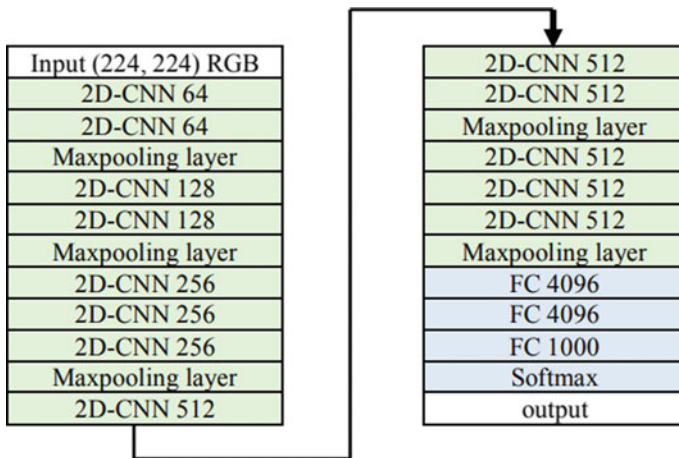
Mini-batch variance:  $\sigma_\beta^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_\beta)^2$

Normalize  $\hat{x}_i \leftarrow \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}}$

Scale and shift:  $y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$

As stated earlier, VGG16 is a convolutional neural network presented by Simonyan and Zisserman [37]. The structure of VGG16 shown in Fig. 5.13 consists of 13 convolutional layers which are worked as feature extraction and three fully connected neural network layers (FC) which are aimed to classify input data. Max pooling layers are also inserted at the end of each convolutional layer group. Moreover, Softmax layer is added at the end of the model to provide a classified result. VGG16 achieves a high rate of recognition performance at 89.3% of mean Average Precision (mAP) on VOC-2012 dataset which is an image dataset of action classification. The original VGG16, which is a 2D CNN model, supports only an image task. Therefore, VGG16 is reinvented to support HAR video dataset for this study.

The experiment is divided into three sections such as doing video preprocessing, building 3D CNN models with and without Batch Normalization, and finally training both models to compare the model accuracy of training and test data. Even though the



**Fig. 5.13** Original VGG16 model

originally spatial dimension of UCF101 is (320, 240, 3) which represents to width, height, and color channel, respectively, for each image frame, it is compressed to (224, 224, 3) to support our models. In the temporal term, it is sampled equally 6 frames for each video clip. Therefore, the input data dimension after being resized is (224, 224, 6, 3) which means width, height, temporal data, and channel, respectively, for each sample. In terms of label data, the dimension is (101) for each sample. This means that there are 101 classes of human activities. UCF101 contains totally 13,320 samples. Therefore, the processed input data and label data dimension are (13,320, 224, 224, 6, 3) and (13,320, 101), respectively. The processed input data and label data are stored and compressed to a single file as .npz data type to provide a convenient process for training and testing a deep learning model. The processed dataset (.npz file) size is 12.4 GB. Finally, the processed dataset is also randomly separated into two groups for training and test data. The ratio between training and test data is 80:20. The video preprocessing method is presented in Fig. 5.14.

There are two models used for this section. They are built by Keras which is one of the most practical frameworks to build deep learning models. Model 1 (3D CNN without BN) as shown in Fig. 5.15 consists of two sections such as feature extraction and classification. The feature extraction is reinvented from VGG16 which is a 2D CNN model for image classification. To support video dataset, 2D convolutional and 2D max pooling layers are replaced by 3D Convolutional and 3D max pooling layers, respectively. For the classification section, there is one layer of neural network added. It consists of 512 nodes and uses Sigmoid as an activation function. Softmax is the last layer added at the end of the model to provide the results of video classification. Model 2 (3D CNN with BN) as shown in Fig. 5.15 is similar to the first model except adding Batch Normalization at the end of each convolutional layer group. There are five convolutional groups such as 3D-CNN-64, 3D-CNN-128, 3D-CNN-256, and two 3D-CNN-512. This means that the input data is normalized before feeding into four convolutional groups (3D-CNN-128, 3D-CNN-256, and two 3D-CNN-512),

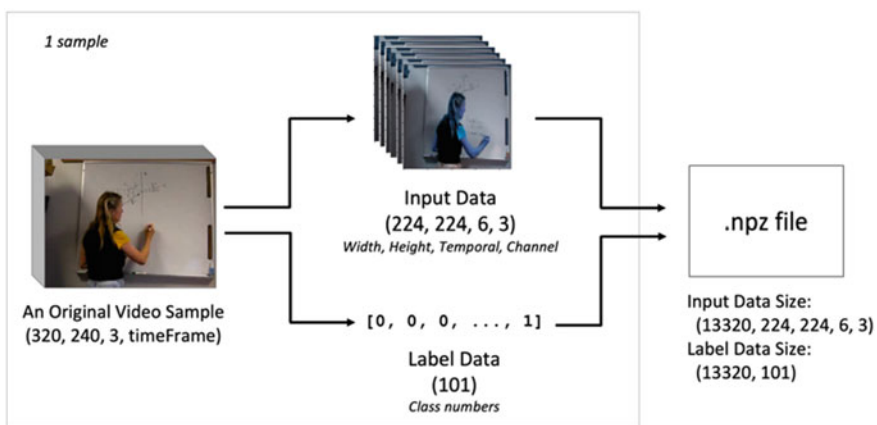


Fig. 5.14 Video preprocessing of UCF101 dataset

| Model 1<br>3D CNN <b>without BN</b>                  | Model 2<br>3D CNN <b>with BN</b>                     |
|--|--|
| 3D-CNN 64  | 3D-CNN 64  |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 64  | 3D-CNN 64  |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D Max Pooling                                       | 3D Max Pooling                                       |
|  | <b>Batch Normalization</b>                           |
| 3D-CNN 128   | 3D-CNN 128   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 128   | 3D-CNN 128   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D Max Pooling                                       | 3D Max Pooling                                       |
|  | <b>Batch Normalization</b>                           |
| 3D-CNN 256   | 3D-CNN 256   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 256   | 3D-CNN 256   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 256   | 3D-CNN 256   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D Max Pooling                                       | 3D Max Pooling                                       |
|  | <b>Batch Normalization</b>                           |
| 3D-CNN 512   | 3D-CNN 512   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 512   | 3D-CNN 512   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 512   | 3D-CNN 512   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D Max Pooling                                       | 3D Max Pooling                                       |
|  | <b>Batch Normalization</b>                           |
| 3D-CNN 512   | 3D-CNN 512   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 512   | 3D-CNN 512   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D-CNN 512   | 3D-CNN 512   |
| Activation(relu)                                     | Activation(relu)                                     |
| 3D Max Pooling                                       | 3D Max Pooling                                       |
|  | <b>Batch Normalization</b>                           |
| Flatten layer  | Flatten layer  |
| Neural Network 512 nodes<br>with Activation(sigmoid) | Neural Network 512 nodes<br>with Activation(sigmoid) |
| Softmax layer  | Softmax layer  |

Fig. 5.15 3D CNN models without BN and with BN

and one neural network. The BN parameters used for this study are that momentum is 0.99 and epsilon is 0.001.

Adding BN increases the numbers of model parameters. Therefore, there are totally 44,450,085 and 44,455,973 parameters for the model without and with BN, respectively. When training the models performed by 16 GB GPU (NVIDIA Tesla V100), Stochastic Gradient Descent is used to run as an optimizer. Two models are trained and tested in every epoch for 100 epochs with the processed data. The parameters for training both models are that batch size is 16, and the learning rate is 0.01. A 3D CNN model with and without BN are compared. Three results are presented in this section such as the results of loss function values and model's accuracy. One of the most useful measurements of successfully training CNN model is a low value of loss function which implies an error of trained model output compared to label data. Figure 5.16 presents loss function values of training and test data, respectively. It is clear that a model with BN provides a lower value than a model without BN at the starting point. A loss function value also drops dramatically when using BN while the value seems unchanged for the first five epochs when using no BN. Even though the loss function value of both models continues to decrease, a model without BN is not able to be trained after epoch 19. It cannot find a lower loss function value; then,

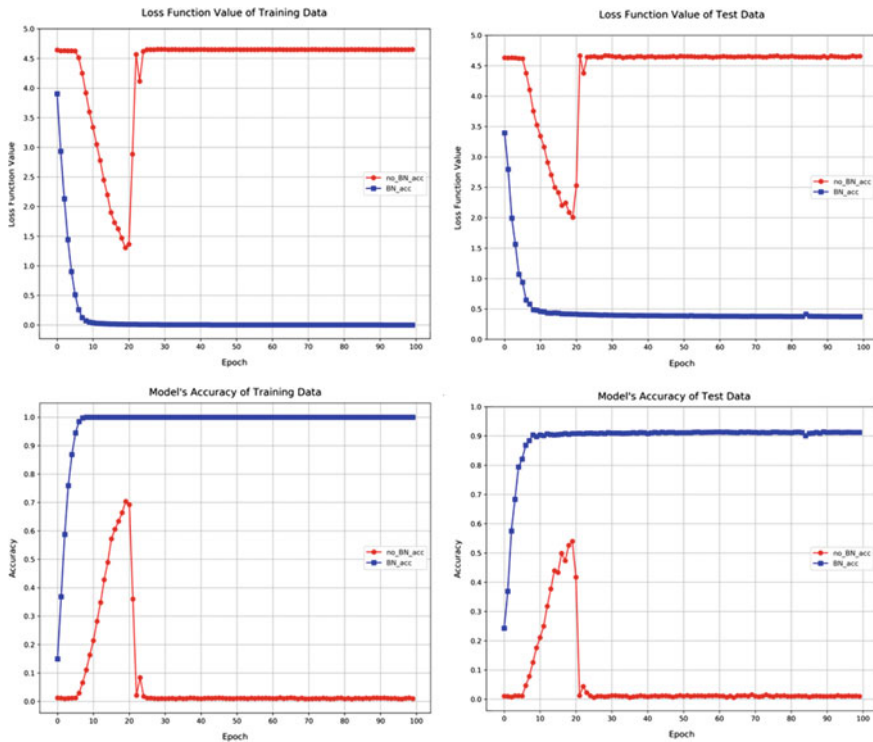
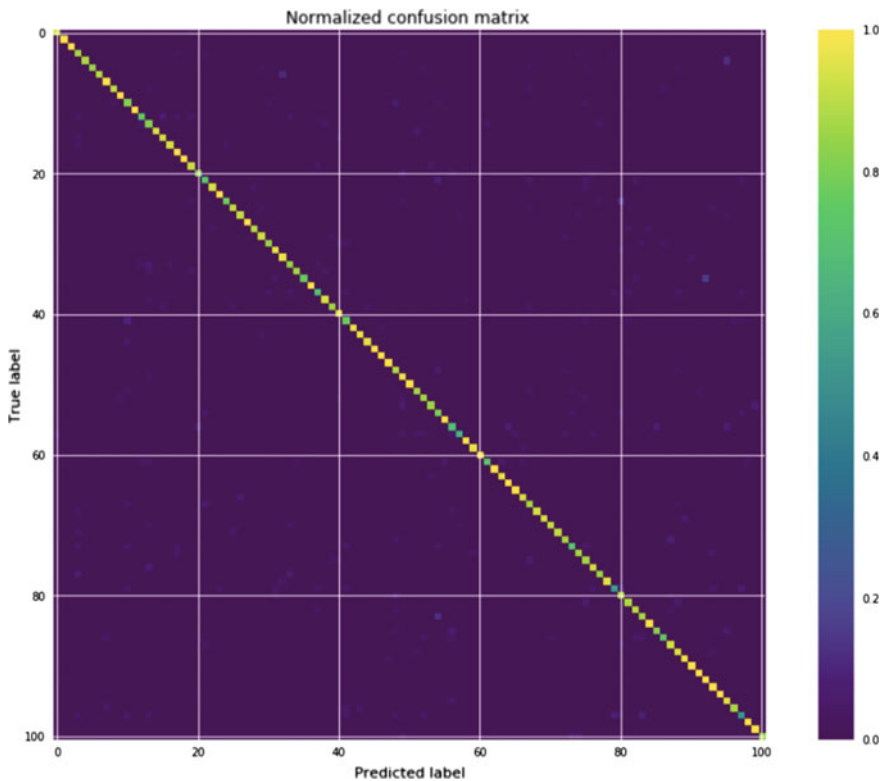


Fig. 5.16 Model training and testing accuracy and loss

the model is trained unsuccessfully. On the other hand, a model with BN is trained successfully and it reaches a loss function value at nearly zero since approximately epoch 10 for the training data. The accuracy of both models is shown in Fig. 5.16 for training and test data, respectively. Furthermore, the confusion matrix is shown in Fig. 5.17. The results show that a model with BN provides a higher rate of model accuracy than a model without BN. The accuracy of a model using BN increases dramatically, and it reaches a saturate point at epoch 10 onward at approximately 100 and 90% for training and test data, respectively.

On the other hand, even though the accuracy of a model without BN increases dramatically since epoch 8, it stops at epoch 19 which reaches its highest accuracy at 70.36 and 54.01% for training and test data. After that, the accuracy drops to nearly zero which means that this model is unreliable to be used for video classification of HAR tasks. To confirming the accuracy of the improved model, the cross-validation method has been performed. According to Fig. 5.16, it finds that the 3D CNN model reinvented from VGG16 with adding BN, which is the improved model, provides high accuracy rate for both training and test data. However, training and testing a



**Fig. 5.17** Confusion matrix for the model with BN trained and tested with UCF10 dataset. Labels are from UCF101 dataset starting from 0 for Apply Eye Makeup and ending to 100 for Yo Yo

**Table 5.5** Cross-validation of the improved model

| Cross-validation group | Accuracy (%)  |           |
|------------------------|---------------|-----------|
|                        | Training data | Test data |
| 1                      | 100           | 90.9      |
| 2                      | 100           | 90.4      |
| 3                      | 100           | 91.2      |
| 4                      | 100           | 92.6      |
| 5                      | 100           | 90.7      |
| Average                | 100           | 91.2      |

deep learning model on different data samples could provide different accuracy even though doing within the same dataset. Therefore, a cross-validation method [33] is performed to confirm model accuracy. For the cross-validation method, UCF101 dataset is divided into five groups and then each data group is trained and tested on the improved model (the 3D CNN model with BN). The result of applying cross-validation is shown in Table 5.5. It found that the improved model can still achieve high accuracy rate. The training and test's mean accuracy are at 100 and 91.17%, respectively. When comparing the accuracy of the improved model to the state-of-the-art models which Joao Carreira and Andrew Zisserman study [8] on a UCF101 benchmark, it found that the improved model can be ranked in a group of state-of-the-art accuracy presented in Table 5.6. There are two main points to present in this experiment. Firstly, this study suggests using BN to improve model accuracy when developing a deep learning model from scratch to make it achieve a high accuracy rate. Two models are reinvented from VGG16. The difference between both models is adding BN and no adding BN. The model with BN provides a higher accuracy rate at the same epoch when comparing to the model without BN. Even though training both models for 100 epochs consumes the same amount of time—estimated 13 h,

**Table 5.6** State-of-the-art models on the UCF101 benchmark

| Model  | Accuracy %  |
|--|-------------|
| Two-Stream I3D, Imagenet + Kinetics pre-training [8] | 98.0        |
| ST-ResNet + IDT [13]                                 | 94.6        |
| Temporal Segment Networks [47]                       | 94.2        |
| Two-Stream Fusion + IDT [14]                         | 93.5        |
| TDD + IDT [46]                                       | 91.5        |
| <b>Our model</b>                                     | <b>91.2</b> |
| C3D ensemble + IDT, Sport 1 M pre-training [41]      | 90.1        |
| Dynamic Image Networks + IDT [5]                     | 89.1        |
| Two-Stream [36]                                      | 88.0        |
| IDT [43]   | 86.4        |

the speed of reaching a saturation point of the model with BN is also faster. In other words, it requires less epoch to train than the model without BN. The model with BN requires 10 epochs to reach the saturation point which consumes training time estimated 1 h and 20 min while the model without BN requires 19 epochs which consumes estimated 2 h and 30 min to train the model. Secondly, 3D CNN model reinvented from VGG16 with BN is considered as an improved model. It consists of 44,455,973 parameters which require 178-MB memory (not include model structure) to store as a .hd5 data format. This model achieves a high accuracy rate of 91.2% which is in the state-of-the-art UCF101 benchmark results.

## 5.4 Conclusion

The results of the first experiment demonstrate that a high accuracy for activity recognition on the complex SVW dataset is achievable. The best technique is a VGG16 model where the testing accuracy 80.7% is obtained. This could be due to the fact that this model consumes more information throughout the video. When zooming in on this best model and looking at other performance measures, it becomes clear that the misclassification can be explained by looking at the frames. Some high scores from the SVW and UCF101 confusion matrices have been investigated to see whether the frames can tell more about these misclassifications. One solution is to apply data preprocessing accordingly or object recognition for similar sports as discussed earlier. In another experiment, we have used the famous UCF101 dataset and trained a CNN using Batch Normalization. It significantly improved the results to 90.4% for 101 different types of sports activities. The models presented in this chapter are suitable to be deployed in smart environments in order to identify different activities. One scenario is to upload a short clip of the activity to the cloud, and then it will be identified as a certain type of sport. This record could be kept for different purposes such as monitoring healthy living standards or training in countries/schools where budget and automatization are important. The methods presented in this chapter rely on off-the-shelf hardware such as ordinary smart phones.

The results of training a CNN model with and without BN in our second experiment on UCF101 dataset found that a model with BN provides a lower loss than another model for both training and test datasets. This results in a higher accuracy rate. The model with BN achieves an accuracy of approximately 100 and 90% for training and test datasets, respectively, while a model without BN reaches its highest accuracy of 70.36 and 54.01% for training and test data. Once the accuracy of a model using BN reaches its highest point, it continues to remain at this level. However, the accuracy of a model without BN drops after it reaches its highest point. This study also observes that a model with BN requires nearly two times fewer epochs to reach its highest model accuracy, compared to a model without BN. Therefore, adding BN is a suggestion to provide a high accuracy rate of a CNN model especially when building from scratch.

In addition to these, it also finds that the improved model (3D CNN reinvented from VGG16 with BN) can achieve a high model accuracy of 91.2% after confirmation by the cross-validation method. This level of accuracy is in the range of the state-of-the-art UCF101 benchmark results. However, the accuracy of training data is higher than test data; this means that the model is still confronting an overfitting problem [4]. There is room for model accuracy improvement in the future, by adding L1 and L2 regularization [22], adding dropout [39], or using more data to train the models.

To finish, this chapter gives an overview of how to utilize different techniques in machine learning to tackle some daily and routine tasks. We showed that using pre-trained models and modifying their parameters before applying to another set of data can produce a very accurate and robust model to identify human activities in different uncontrolled environments. One application of this model can be to identify sports or activity types in a smart home/gym platform and record the data. This data can be used to enhance these activities in smart homes over a period of time. It is also beneficial to modernizing smart sports hall to detect and record different sports automatically for an audit, for example.

## References

1. Aggarwal J, Cai Q (1999) Human motion analysis: a review. *Comput Vis Image Underst* 73:428–440
2. Aggarwal JK, Ryoo M (2011) Human activity analysis: a review. *ACM Comput Surv (CSUR)* 43:1–43
3. Asperger J, Poore A (2017) Convolutional neural networks for classification of noisy sports videos. s.l., s.n.
4. Bilbao I, Bilbao J (2017) Overfitting problem and the over-training in the era of data: particularly for artificial neural networks. s.l., s.n., pp 173–177
5. Bilen H et al (2016) Dynamic image networks for action recognition. s.l., s.n., pp 3034–3042
6. Bishop C (2006) *Pattern recognition and machine learning*. NJ Springer, Secaucus
7. Blank M (2007) Actions as space-time shapes. *Trans Pattern Anal Mach Intell* 29:2247–2253
8. Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. s.l., s.n., pp 6299–6308
9. Chen L, Hoey J, Nugent CD, Cook DJ, Yu Z (2012) Sensor-based activity recognition. *IEEE Trans Syst Man Cybern Part C (Appl Rev)* 42(6):790–808
10. Chen L et al (2017) Why batch normalization works? A buckling perspective. s.l., s.n., pp 1184–1189
11. Chen X (2017) Deep manifold learning combined with convolutional neural networks for action recognition. *IEEE Trans Neural Netw Learn Syst* 99:1–15
12. Dalal N, Triggs B, Schmid C (2016) Human detection using oriented histograms of flow and appearance. s.l., s.n., pp 428–441
13. Feichtenhofer C, Pinz A, Wildes R (2016) Spatiotemporal residual networks for video action recognition. s.l., s.n., pp 3468–3476
14. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. s.l., s.n., pp 1933–1941
15. Gavrilă D (1999) The visual analysis of human movement: a survey. *Comput Vis Image Underst* 73:82–98
16. Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT Press, s.l.



17. Hongeng S, Nevatia R, Bremond F (2004) Video-based event recognition: activity representation and probabilistic recognition methods. *Comput Vis Image Underst* 96(2):129–162
18. Ikizler N, Duygulu P (2007) Human action recognition using distribution of oriented rectangular patches. In: *Proceedings of the 2nd conference on human motion: understanding modelling, capture and animation*, vol 2, pp 271–284
19. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*
20. Jaimes A, Sebe N (2005) Multimodal human computer interaction: a survey. *ICCV-HCI*, s.l.
21. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, vol 25
22. Kukačka J, Golkov V, Cremers D (2017) Regularization for deep learning: a taxonomy. *arXiv preprint arXiv:1710.10686*
23. Lin Q, Zhang D, Chen L, Ni H, Zhou X (2014) Managing elders' wandering behavior using sensors-based solutions: a survey. *Int J Gerontol* 2:49–55
24. Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: *CVPR*, pp 3337–3344
25. Lowe D (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60:91–110
26. Marszałek M, Laptev I, Schmid C (2009) Actions in context. *s.l., s.n.*, pp 2929–2936
27. Ni B, Wang G, Moulin P (2011) Rgb-d-hudaact: a color-depth video database for human daily activity recognition. In: *IEEE international conference on computer vision workshops (ICCV workshops)*, s.l.
28. Ojala T, Pietikainen M, Maenpää T (2002) Multiresolution gray scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24:971–987
29. Pantic M, Pentland A, Nijholt A, Huang T (2007) Human computing and machine understanding of human behavior: a survey. *Lect Notes Artif Intell* 4451:47–71
30. Poppe R (2010) A survey on vision-based human action recognition. *Image Vis Comput* 28:976–990
31. Rachmadi RF, Koutaki G, Uchimura F (2016) Combined convolutional neural network for event recognition. In: *The Korea-Japan joint workshop on Frontiers of Computer Vision (FCV)*
32. Rachmadi RF, Koutaki G, Uchimura N (2016) Video classification using compacted dataset based on selected key frames. In: *IEEE region 10 conference (TENCON)*
33. Refaailzadeh P, Tang L, Liu H (2009) Cross-validation. In: *Encyclopedia of database systems*, pp 532–538
34. Safdarnejad SM (2015) Sport videos in the wild (SVW): a video dataset for sports analysis. In: *Proceedings of international conference on automatic face and gesture recognition*
35. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. *s.l., s.n.*
36. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: *Computer vision and pattern recognition*
37. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *ICLR*
38. Soomro K, Zamir AR, Shah M (2012) UFC101: a dataset of 101 human action classes from videos in the wild. *CRCV-TR-12-01*
39. Srivastava N et al (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15:1929–1958
40. Szegedy C, Vanhoucke V, Loffe S, Shlens J (2015) Rethinking the inception architecture for computer vision
41. Tran D et al (2015) Learning spatiotemporal features with 3d convolutional networks. *s.l., s.n.*, pp 4489–4497
42. Vrigkas M, Nikou C, Kakadiaris L (2015) A review of human activity recognition methods. *Front Robot AI* 2:1–28
43. Wang H, Schmid C (2013) Action recognition with improved trajectories. *s.l., s.n.*, pp 3551–3558

44. Wang J, Chen Y, Hao S, Peng X, Hu L (2019) Deep learning for sensor-based activity recognition: a survey. *Pattern Recognit Lett* 3–11
45. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: *IEEE conference on computer vision and pattern recognition*, pp 1290–1297
46. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. *s.l., s.n.*, pp 4305–4314
47. Wang L et al (2016) Temporal segment networks: towards good practices for deep action recognition. *s.l., s.n.*, pp 20–36
48. Wrzalik M, Krechel D (2017) Human action recognition using optical flow and convolutional neural networks. In: *16th IEEE international conference on machine learning and applications*, pp 801–805
49. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden Markov model. In: *Proceedings of IEEE conference on computer vision and pattern recognition*, pp 379–385

# Chapter 6

## Object Detection-Based Location and Activity Classification from Egocentric Videos: A Systematic Analysis



**Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas P. J. J. Noldus and Remco C. Veltkamp**

**Abstract** Egocentric vision has emerged in the daily practice of application domains such as lifelogging, activity monitoring, robot navigation and the analysis of social interactions. Plenty of research focuses on location detection and activity recognition, with applications in the area of Ambient Assisted Living. The basis of this work is the idea that indoor locations and daily activities can be characterized by the presence of specific objects. Objects can be obtained either from laborious human annotations or automatically, using vision-based detectors. We perform a study regarding the use of object detections as input for location and activity classification and analyze the influence of various detection parameters. We compare our detections against manually provided object labels and show that location classification is affected by detection quality and quantity. Utilization of the temporal structure in object detections mitigates the consequences of noisy ones. Moreover, we determine that the recognition of activities is related to the presence of specific objects and that the lack of explicit associations between certain activities and objects hurts classification performance for these activities. Finally, we discuss the outcomes of each task and our method's potential for real-world applications.

**Keywords** Egocentric vision · Object detection · Location classification · Activity classification · Detection quality · Temporal associations

---

Parts of this chapter are © 2018 IEEE. Reprinted, with permission, from Kapidis et al. [1].

---

G. Kapidis (✉) · E. van Dam · L. P. J. J. Noldus  
Noldus Information Technology, Wageningen, The Netherlands  
e-mail: [georgios.kapidis@noldus.nl](mailto:georgios.kapidis@noldus.nl); [g.kapidis@uu.nl](mailto:g.kapidis@uu.nl)

G. Kapidis · R. Poppe · R. C. Veltkamp  
Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

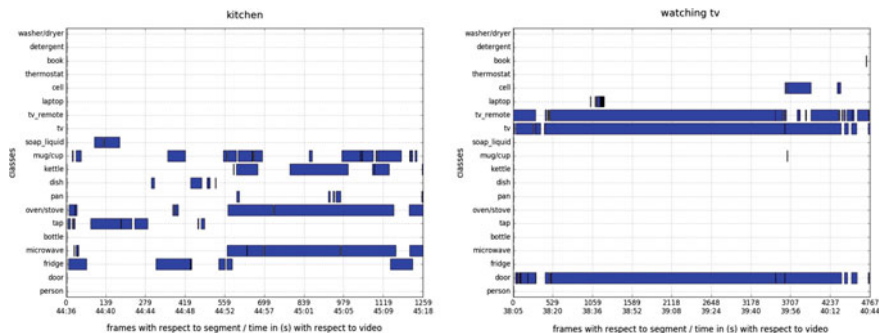
## 6.1 Introduction

Egocentric vision is an essential part of computer vision with applications in conventional fields such as activity recognition [2] and video summarization [3] as well as in more elaborate, for instance social interaction analysis [4], guideline generation for visual assistance [5] and infant visual attention [6]. In this work, we focus on indoor location and activity detection from egocentric videos, with typical applications in Ambient Assisted Living (AAL) [7]. An example can be non-intrusive status updates to healthcare professionals about the locations and actions of people suffering from limited vision or dementia. Activities of daily living are also of interest when it comes to patient rehabilitation after a serious illness. Normally, this process would take place in a protected environment, far from the person's home. The possibility of continuous and real-time monitoring offered by egocentric cameras allows for noninvasive and personalized care. Reusability of the equipment by other patients at the end of the recovery period is an additional incentive toward adoption by nursing homes. Moreover, enhancement with intelligent detection mechanisms will promote privacy, since only information relevant to the rehabilitation will need to be communicated to third parties and not the actual video stream. An example use-case is that of dementia patients who require constant monitoring and professional care [8]. Egocentric vision is able to provide the indoor location [1], the duration of physical exercises [9] or the performed activity, upon request or in a continuous mode.

The use of egocentric cameras is alluring as they are becoming smaller and less intrusive, two essential qualities for wearables targeting everyday use. They can be used as an alternative to expensive multi-sensor installations that convert an existing house into a smart-home. By taking advantage of recent advances in machine learning, a single sensor—the egocentric camera—will produce information about the location in the house, sociability or loneliness, performed activity and even imminent dangers stemming from the latter.

To produce an inference on an image or video frame, one could calculate image-descriptive features [10, 11] stack them in vectors and classify, using machine learning models in a supervised fashion. In recent years, feature extraction and classification have merged into end-to-end deep networks, providing promising results. In this work, we take a step back and consider a different type of input.

Our key idea is to use the detected objects in a video frame as cues to recognize the indoor location or an ongoing activity. Initially, we build on the idea that rooms can be characterized by the presence of specific, distinctive objects. This consistency can be translated into associations between objects and locations. Consider, for example, Fig. 6.1 (left) which shows the detected objects of an egocentric video segment from a kitchen. If we categorize the objects based on their mobility, we may group them into (a) those that can be thought of as *movable*, but bear meaning for understanding the scene, such as the soap, the mug and the dish and (b) those that are *unmovable* but (i) distinctive to this particular location, such as the stove, the microwave or the fridge and (ii) those that can be found in multiple locations, for example the tap, which could also appear in a bathroom. Similarly, we claim that the activity of the



**Fig. 6.1** Detected objects for ‘kitchen’ (left). Certain objects, such as the oven and the microwave, dominate the scene. For activity ‘watching TV’ (right), the television and the remote controller are indicative of the performed activity

egocentric protagonist can be inferred from the detected objects, considering Fig. 6.1 (right) which shows a TV and a remote controller for most of the duration of a video segment for activity ‘watching TV.’

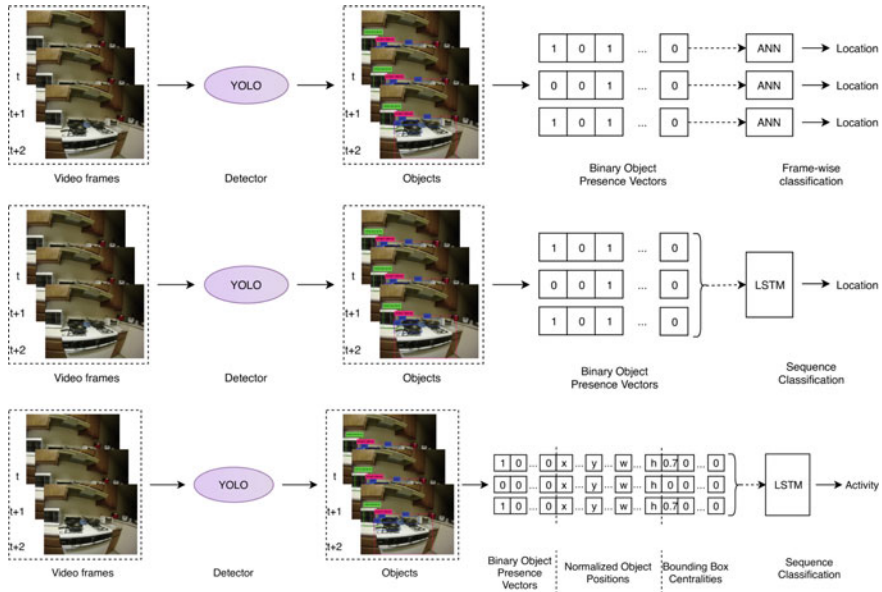
This motivates us to perform an analysis on the videos of the Activities of Daily Living (ADL) dataset [12] to discover associations between objects and locations and objects and activities. For the object-location associations, we train classifiers with artificial neural networks (ANNs) and long short-term memory (LSTM) networks [13] to experiment with per frame classification and utilization of the temporal structure of the data, respectively. Conceptually, an individual frame of a scene might include only partial information about the objects, as not all that are detectable may fit in view. However, the combination of multiple frames over time can encode a more complete view of the room. Eventually, we compare the performance of classifiers from both types of models, trained either on object labels or detections, from detectors trained on object categories from different datasets.

For the object-activity associations, we rely on detections enhanced with certain appearance features. Apart from the presence of objects, we measure the bounding box sizes and positions in the frame. We aim to investigate whether this additional information modifies the status of an object as participating in the ongoing activity, for example, when observed from a distance (smaller) or at the edge of the view. Figure 6.2 outlines our approaches.

The *contributions* of this work<sup>1</sup> can be summarized to:

- The development of a method to analyze object associations toward (1) locations and (2) activities in egocentric videos,
- The object presence feature, which despite its simplicity demonstrates acceptable performance,

<sup>1</sup>Code and data for our experiments are located here: <https://github.com/georkap/object-based-location-classification>.



**Fig. 6.2** Our pipelines for location and activity recognition. We extract objects from video frames using an object detector (YOLO [14]). The detections for a frame are turned into binary object presence vectors. This representation is the input for classification with ANNs (top) which produce one location per frame. The presence vectors are stacked in sequences and used as input to LSTMs (middle) which generate one location prediction for the complete sequence. For activity classification (bottom), we consider two additional features that describe the bounding boxes and classify them with LSTM

- The description of location classification results for diverse object sets and detection thresholds with and without temporal information,
- The demonstration that laborious object annotations are not required for location classification, given that our system performs equally well using only automatic detections and
- The analysis of object-activity associations in the context of daily living and the effect that object sizes and positions have in the activity recognition results.

Section 6.2 is an overview of related work in egocentric object, location and activity recognition, datasets and applications in the field. In Sect. 6.3, we describe the dataset we used, the object detectors and the methodology for both tasks. In Sect. 6.4, we present our results. Our findings are discussed in Sect. 6.5, and Sect. 6.6 concludes this chapter.

## 6.2 Related Work

### 6.2.1 *Objects and Location Classification*

We focus on recognizing indoor locations based on the detected objects from egocentric videos of people moving freely in their homes. The ADL dataset [12] has these characteristics and the required object annotations. The scientific literature provides a plethora of egocentric datasets [15–21]. The datasets of [15, 16] are created with the aim of detecting locations and observing indoor and outdoor everyday scenes. The dataset of [21] focuses on activities that take place either indoors or outdoors, such as walking, running and sitting, whereas [17] is enhanced with accelerometer and heart rate data to infer the level of sedentariness in performed activities. The dataset of [18] includes annotations and segmentations of the important objects that characterize the activities, with a large number of the videos being outdoors. Datasets from [19, 20] consist of videos in a kitchen, in which the participants are asked to prepare food according to predefined recipes.

Understanding of locations, in terms of mapping the surrounding area with image features or semantically labeling the environment, is actively under research in egocentric vision. In [22], a combination of scene illumination and distinct location characteristics is learned in an unsupervised way, in order to enhance the usability of wearable cameras for hand detection. Location recognition is indirectly the task in [23] where a Google Glass application captures images of the user’s field of view and retrieves information about the buildings in sight. An indoor localization system is considered in [24] where the combination of a camera and a 2D laser scanner is applied to register query images from users into a real-world coordinate system. A multi-view indoor localization system based on image features is proposed in [25]. It distinguishes the indoor locations by computing self-similarity matrices from the extracted images to correlate the various captured views of the scene. Afterward, it learns the equation system through these features and when a query image is given, it can provide its location and orientation. The combination of wearable egocentric stereo cameras and inertial sensors is considered in [26] to map an outdoor workspace and provide routing guidance for executing specific tasks in the workplace. Another system for location classification is described in [27]. Visual recognition is based on low-level features and semi-supervised training procedures to take advantage of sparsely annotated available data. Instead, we use high-level features, the object detections from every frame. We do not depend on previous knowledge of the specific locations that the users find themselves into but build our inference upon characteristic objects that are detected in them.

Temporal segmentation of egocentric videos is considered in [15] to highlight personal locations of interest. Training is based on user-provided frame samples of locations labeling those capturing user interest as positive. Then, the system learns to reject the frames that do not depict locations relevant to the user. Personal locations are also analyzed in [16] as part of a user’s daily routine. Classification of the video frames into locations relies on either convolutional neural network (CNN) features

or handcrafted ones. In [28, 29], the combined improvement of object detection and scene identification is investigated. Initially, scene identification is performed based on temporally associated CNN features and its output is used to improve the results of object detectors by linking the objects to specific locations. Eventually, they show that by using an LSTM to train on the temporal sequences of the detected objects directly, it is possible to improve the performance of detectors without explicitly using extracted features about locations. They perform their experiments on the ADL dataset. Our work is the opposite of this concept, where we use the object detections to infer either locations or activities.

Searching through the parameters of a model to find the optimal configuration is a common theme in machine learning [30–32]. In [30], various LSTM variants are tested on speech recognition, handwriting recognition and music modeling tasks to inspect the differences introduced by the changing network architectures. It is shown that most LSTM variants do not improve significantly, if at all, over the default LSTM structure, which performs relatively well for all considered tasks. Variations in the hyperparameters used for training are also explored, but it is observed that they are uncorrelated. In an effort to provide further insights into the reasons behind the effectiveness of LSTM, [31] provide a thorough study on cell activations, error analysis and data representations. A more recent approach to tuning a model’s hyperparameters is presented in [32] where effort is put into balancing regularization for a particular dataset and the respective architecture utilized for training, by studying the loss during both training and testing for signs of over- and underfitting. Our aim is not to compare the parameters that define the structure of networks, such as the number of layers or neurons per layer. Rather, we vary the model from ANN to LSTM to augment the learning process with temporal information. In the context of searching for the optimal parametrization for location classification, our work comprises a large-scale search over the dataset combinations for training and testing the classifiers, the object detectors that generate the detection datasets and the effect of the detection confidence in detection quality and quantity.

### ***6.2.2 Objects and Activity Classification***

Human action recognition from video is a computer vision task that introduces multiple challenges to researchers, ranging from the innate difficulties of video analysis, such as illumination changes and motion blur, to the adversities of activity recognition, including class variability, viewpoint variation and scarcity of training data [33].

In egocentric activity recognition, the person is removed from view; therefore, the inference must rely on indirect cues. These include the detection of hands, relevant objects and locations [2, 34], as well as convolutional features [35], image features [36] or motion-based features, such as ego-motion and global motion or combinations [16, 20, 36–38]. In [34], the hands are used for activity recognition. An egocentric hand detector is described, where region proposals are combined with



convolutional neural networks (CNNs) for hand classification. The correct instances are turned into pixel-level segmentations, which in turn are the input in a second CNN classifier to infer the performed activity. In [35], objects and activities are analyzed in close relation, with each affecting the decision for the other interchangeably. Objects in association with hands and the interactions between them are modeled to infer activities and their associations in [38]. We take a more elementary approach and study only the objects and their characteristics for activities, without feedback between modalities.

Motion-based works usually rely on optical flow as input to train machine learning models. In [37], each frame is divided into a set of grid cells and a single sparse optical flow value is extracted for each cell per frame. The flow values are used as train data to a CNN and classified into specific activity classes. In [36], scene, object, hand and head movements are modeled with dense trajectories, color histograms and local binary patterns and used as inputs to support vector machines for the classification of the food-related activities of the GTEA datasets [20]. The aim is to measure the individual effect of new features on activity classification performance by adding them gradually, based on the idea that their contributions are complementary. We also follow this concept of feature aggregation.

Combination of detections along with motion cues for the description of activities is not uncommon. In [2], multiple networks are utilized for the extraction of hand poses and object segmentations per frame. At the same time, a different network is trained on optical flow to predict short-term actions. Eventually, all networks are combined toward a new output, the activity prediction. Our method is different in that the object detection module is fully detached from the activity prediction and can be thought of as a different component that could be substituted with a more efficient one, when it becomes available. In [39], videos are mapped onto a semantic graph, with nodes for each freely annotated object and action, trained on the visual similarities between them. The activities in unseen videos are recognized as probability distributions among the existing action labels. They demonstrate the inherent interactions that occur between objects and actions, an idea that relates to our work.

Activity recognition solely based on the detected objects in the scene is considered in [40, 41]. In [40], object detection relies on video input, complemented with RFID tags, manually placed in the scene. Our work only considers video. In [41], a dynamic feature prioritization policy is developed to choose which single-class object detectors to promote, in an effort to execute as few of them as possible, thus saving computations, while also maximizing the classification accuracy on the subsequent frame. The aim is to take advantage of the spatiotemporal correlations that occur during an activity and avoid extraction of unnecessary features, which would not provide important information for the recognition process as a whole. Our method does not focus on single objects and their possible associations through time during detection, but takes advantage of the state of the art in single-frame, multi-class object detection to extract objects in real time and uses LSTM to learn the temporal associations.

## 6.3 Methodology

In this section, we analyze an egocentric video dataset in terms of objects, locations and activities (Sect. 6.3.1) and select the object detection framework to facilitate our tests (Sect. 6.3.2). Moreover, we study and discuss the parameters of the location (Sect. 6.3.3) and activity (Sect. 6.3.4) classification tasks.

### 6.3.1 *Activities of Daily Living (ADL) Dataset*

The ADL dataset [12] consists of 20 videos of people performing activities occurring indoors, captured from the egocentric perspective. Each video is a record of the subject's choice of activities from a predefined set, performed in an unscripted manner. In every video, the subject is different and operates in their own house, thus providing considerable variations in locations and activities, among videos. In total, there are approximately 10 h of egocentric videos, equivalent to more than one million frames. The videos are annotated with activity labels with start/end times, object bounding boxes, object tracks and human-object interactions. Train and test splits are provided by the authors; videos 1–6 are considered training data, and the remaining 14 comprise the test set. For our experiments, we use the same splits.

Originally, in [12], there are annotations for 48 object classes, but due to the low number of either training or testing samples, only 42 are considered for their tests. For our object detectors, we use either the whole set of 48 classes or a subset of 20. We elaborate on object detection in Sect. 6.3.2. A list of the object classes together with their occurrences in the ADL dataset appears in Table 6.1.

In Sect. 6.3.3, we are interested in the analysis of locations, so we extend the dataset with the location annotations from [29]. For every 30 frames of video, one location class out of the eight possible is annotated, namely kitchen, bedroom, bathroom, living room, laundry room, corridor, outdoor and undefined (Table 6.2). Class 'undefined' occurs in blurred frames or non-identifiable locations. We do not use these frames for training and testing the location classification models. Hence, the location classes are seven in our experiments.

In Sect. 6.3.4, we focus on activity classification and make use of the existing activity annotations in the ADL dataset. We transform the labels from describing video segments with specific start and end times, to one activity per frame. The activities are shown in Table 6.3. In [12], only 18 activities are considered whereas the dataset contains labels for 33. We consider all 33 activities in our experiments.

**Table 6.1** Forty-eight object classes of the ADL dataset and the number of occurrences per class. In the third column the instances in the train set. In bold, the classes of the ADL20 subset

| Class name         | Total  | Train |
|--------------------|--------|-------|
| <b>Person</b>      | 4650   | 2424  |
| <b>Door</b>        | 7903   | 2019  |
| <b>Fridge</b>      | 1999   | 301   |
| <b>Microwave</b>   | 2369   | 527   |
| <b>Bottle</b>      | 10,310 | 1705  |
| <b>Tap</b>         | 7826   | 3252  |
| <b>Oven/stove</b>  | 3196   | 1007  |
| <b>Pan</b>         | 3156   | 1026  |
| Trash can          | 2075   | 486   |
| <b>Dish</b>        | 8216   | 2274  |
| Cloth              | 3077   | 78    |
| Knife/spoon/fork   | 4843   | 1893  |
| Food/snack         | 3876   | 741   |
| <b>Kettle</b>      | 1239   | 464   |
| <b>Mug/cup</b>     | 11,050 | 2766  |
| <b>Soap liquid</b> | 8375   | 2658  |
| Pills              | 394    | 148   |
| Basket             | 1588   | 35    |
| Towel              | 4480   | 1961  |
| Toothbrush         | 1795   | 819   |
| Toothpaste         | 1746   | 492   |
| Electric keys      | 1570   | 417   |
| <b>TV</b>          | 5600   | 2033  |
| <b>Remote</b>      | 2813   | 1253  |
| Container          | 5685   | 3821  |
| Shoes              | 3248   | 735   |
| Tea bag            | 359    | 177   |
| <b>Laptop</b>      | 7027   | 2183  |
| Cell phone         | 653    | 271   |
| <b>Cell</b>        | 571    | 238   |
| <b>Thermostat</b>  | 332    | 137   |
| <b>Book</b>        | 4770   | 445   |
| Dental floss       | 547    | 385   |
| Vacuum             | 519    | 116   |
| Electric keys 2    | 118    | 118   |
| Pitcher            | 1208   | 277   |
| <b>Detergent</b>   | 1105   | 297   |

(continued)

**Table 6.1** (continued)

| Class name          | Total | Train |
|---------------------|-------|-------|
| <b>Washer/dryer</b> | 3362  | 954   |
| Bed                 | 783   | 228   |
| Large container     | 558   | 6     |
| Monitor             | 316   | 287   |
| Keyboard            | 107   | 102   |
| Shoe                | 694   | 300   |
| Blanket             | 85    | 31    |
| Comb                | 307   | 51    |
| Perfume             | 550   | 0     |
| Milk/juice          | 366   | 0     |
| Mop                 | 403   | 0     |

**Table 6.2** Sampled frames per location. Class 'undefined' is not used for training and testing

| Location  | Kitchen | Bedroom | Bathroom | Living room | Laundry room | Corridor | Outdoor | Undefined |
|-----------|---------|---------|----------|-------------|--------------|----------|---------|-----------|
| Train set | 3414    | 1821    | 2307     | 2606        | 815          | 45       | 143     | 492       |
| Test set  | 6850    | 3966    | 2285     | 5045        | 1097         | 133      | 906     | 737       |
| Total     | 10,264  | 2285    | 4592     | 7651        | 1912         | 178      | 1049    | 1229      |

**Table 6.3** Thirty-two activity classes in the ADL dataset, plus the background class (35,906, 85,801). In parentheses the number of train and test frames per class, respectively

|   |  |  |                                      |
|---|--|--|--------------------------------------|
| 1: Combing hair (3539, 6267)                | 2: Makeup (8363, 3926)                   | 3: Brushing teeth (27,729, 26,117)       | 4: Dental floss (8543, 2127)         |
| 5: Washing hands/face (15,050, 17,270)      | 6: Drying hands/face (4014, 6743)        | 7: Entering/leaving room (0, 0)          | 8: Adjusting thermostat (1110, 2459) |
| 9: Laundry (28,812, 46,101)                 | 10: Washing dishes (21,249, 45,807)      | 11: Moving dishes (9984, 0)              | 12: Making tea (15,679, 27,265)      |
| 13: Making coffee (6774, 18,974)            | 14: Drinking water/bottle (6565, 12,328) | 15: Drinking water/tap (0, 540)          | 16: Making hot food (8872, 38,619)   |
| 17: Making cold food/snack (14,268, 11,546) | 18: Eating food/snack (6686, 32,180)     | 19: Mopping in kitchen (1020, 8933)      | 20: Vacuuming (3657, 9864)           |
| 21: Taking pills (3237, 4409)               | 22: Watching TV (37,769, 78,086)         | 23: Using computer (20,445, 57,125)      | 24: Using cell (5817, 10,435)        |
| 25: Making bed (0, 6055)                    | 26: Cleaning house (11,360, 12,655)      | 27: Reading book (20,350, 18,016)        | 28: Using mouth wash (420, 570)      |
| 29: Writing (0, 3628)                       | 30: Putting on shoes (5668, 450)         | 31: Drinking coffee/tea (15,226, 33,778) | 32: Grabbing tap water (599, 1170)   |

### 6.3.2 Object Detection

For our object detection experiments, we use the Darknet framework.<sup>2</sup> Our detector is YOLOv2 [14, 42], a real-time object detection system that can operate on input images of various sizes. YOLOv2 is based on the Darknet-19 architecture [14] and consists of 19 convolutional and 5 max-pooling layers. It is pretrained on ImageNet [43] for 1000 classes, for 160 epochs. From this pretrained model, we develop three separate detectors, one for every object dataset we consider.

Our first YOLOv2-based detector is fine-tuned on the 80 classes of the MS COCO dataset [44], and the weights are provided by the authors of [14]. We call this detector ‘COCO’ for short. We train two additional models with this architecture for the object classes of the ADL dataset: (1) ‘ADL48’, on all the classes in Table 6.1 and (2) ‘ADL20’, on the 20 in bold. The selection of classes for ‘ADL20’ follows [29], where they select only classes for which their detector achieves more than 5% average precision (AP).

The reason for the diversification of detectors is that MS COCO and ADL consist of different sets of classes. ADL comprises objects found in homes (Table 6.1), whereas MS COCO is more generic in its categories (Table 6.4). The split between ‘ADL20’ and ‘ADL48’ is an attempt to produce a detector focused on classes with more samples in the training dataset, thus excluding harder to detect classes. We expect this to reduce the classification loss during training and lead to an improved bounding box classifier for the subset.

For both ‘ADL20’ and ‘ADL48,’ we fine-tune the ImageNet weights for 35 k iterations (i.e., batches). During training, we vary the input dimensions of the detectors to learn objects of various sizes. Training hyperparameters are the same as in [14]. The ‘ADL20’ detector achieves 29.84% mean average precision (mAP) and the ‘ADL48’ 11.15%. In Table 6.5, we report the average precision per class for our detectors. They suggest that YOLOv2 creates a more successful detector for the majority of object classes of the ADL dataset than fast R-CNN [45] in [29].

### 6.3.3 Locations

We model the relationship between the objects in a frame or a series of frames to recognize the location. Applying object detection on the videos of the ADL dataset leads to a binary presence vector (BPV) of zeros and ones, for every video frame, with length equal to the number of output classes of a detector, i.e., 80 for ‘COCO,’ 48 for ‘ADL48’ and 20 for ‘ADL20.’ In BPV, we only consider whether an object exists in a scene or not, regardless of the times it may be found. We also experimented with keeping the counts of multiple detections of the same object in a frame using a multiple presence vector (MPV), but without consistent improvements. Location

---

<sup>2</sup><https://pjreddie.com/darknet/>.

**Table 6.4** MS COCO [44] object classes

|              |           |               |            |              |                |            |            |               |               |
|--------------|-----------|---------------|------------|--------------|----------------|------------|------------|---------------|---------------|
| Person       | Bicycle   | Car           | Motorcycle | Airplane     | Bus            | Train      | Truck      | Boat          | Traffic light |
| Fire hydrant | Stop sign | Parking meter | Bench      | Bird         | Cat            | Dog        | Horse      | Sheep         | Cow           |
| Elephant     | Bear      | Zebra         | Giraffe    | Backpack     | umbrella       | Handbag    | Tie        | Suitcase      | Frisbee       |
| Skis         | Snowboard | sports ball   | Kite       | Baseball bat | Baseball glove | Skateboard | surfboard  | Tennis racket | Bottle        |
| Wineglass    | Cup       | Fork          | Knife      | Spoon        | Bowl           | Banana     | Apple      | Sandwich      | Orange        |
| Broccoli     | Carrot    | Hot dog       | Pizza      | Donut        | Cake           | Chair      | Sofa       | Potted plant  | Bed           |
| Dining table | Toilet    | TV            | Laptop     | Mouse        | Remote         | Keyboard   | Cell phone | Microwave     | Oven          |
| Toaster      | Sink      | Refrigerator  | Book       | Clock        | Vase           | Scissors   | Teddy bear | Hair drier    | Toothbrush    |

**Table 6.5** Average precision (%) of ADL20/48 object detectors per class, trained with YOLOv2. Certain classes are particularly challenging. In bold the classes that improve in the reduced dataset. Comparison with [29] using fast R-CNN for the 20-class subset of the ADL dataset

| Object classes [12] | ADL20        | ADL48        | [29]         |
|---------------------|--------------|--------------|--------------|
| Person              | <b>69.0</b>  | 59.49        | 25.74        |
| Door                | <b>23.2</b>  | 17.72        | 5.59         |
| Fridge              | <b>22.75</b> | 12.85        | 24.95        |
| Microwave           | <b>37.8</b>  | 24.81        | 32.35        |
| Bottle              | <b>10.02</b> | 4.59         | <b>11.28</b> |
| Tap                 | <b>59.27</b> | 51.18        | 39.55        |
| Oven/stove          | <b>44.48</b> | 28.15        | 43.02        |
| Pan                 | <b>16.88</b> | 12.46        | 10.99        |
| Trash can           | –            | 9.61         |              |
| Dish                | <b>14.21</b> | 6.22         | 11.19        |
| Cloth               | –            | 4.55         |              |
| Knife/spoon/fork    | –            | 4.8          |              |
| Food/snack          | –            | 9.65         |              |
| Kettle              | <b>22.54</b> | 8.68         | <b>23.83</b> |
| Mug/cup             | <b>15.92</b> | 15.69        | 13.24        |
| Soap liquid         | 21.76        | <b>31.59</b> | 18.77        |
| Pills               | –            | 0.14         |              |
| Basket              | –            | 0.0          |              |
| Towel               | –            | 9.38         |              |
| Toothbrush          | –            | 9.78         |              |
| Toothpaste          | –            | 11.67        |              |
| Electric keys       | –            | 0.73         |              |
| TV                  | <b>52.07</b> | 49.57        | <b>57.58</b> |
| Remote              | <b>52.49</b> | 30.36        | 32.88        |
| Container           | –            | 5.25         |              |
| Shoes               | –            | 0.72         |              |
| Tea bag             | –            | 0.68         |              |
| Laptop              | <b>44.4</b>  | 41.04        | 37.46        |
| Cell phone          | –            | 10.91        | 8.65         |
| Cell                | <b>0.89</b>  | 0.65         |              |
| Thermostat          | <b>24.88</b> | 3.89         | 9.01         |
| Book                | 16.39        | <b>18.04</b> | 12.83        |
| Dental floss        | –            | 0.92         |              |
| Vacuum              | –            | 0.66         |              |
| Electric keys 2     | –            | 0.0          |              |
| Pitcher             | –            | 3.13         |              |
| Detergent           | <b>9.9</b>   | 9.76         | 9.13         |

(continued)

**Table 6.5** (continued)

| Object classes [12] | ADL20        | ADL48 | [29]         |
|---------------------|--------------|-------|--------------|
| Washer/dryer        | <b>37.96</b> | 25.58 | <b>38.86</b> |
| Bed                 | –            | 0.21  |              |
| Large container     | –            | 0.0   |              |
| Monitor             | –            | 0.0   |              |
| Keyboard            | –            | 0.0   |              |
| Shoe                | –            | 0.15  |              |
| Blanket             | –            | 0.0   |              |
| Comb                | –            | 0.1   |              |
| Perfume             | –            | 0.0   |              |
| Milk/juice          | –            | 0.0   |              |
| Mop                 | –            | 0.0   |              |
| Total (mAP)         | 29.84        | 11.15 |              |

labels exist once every 30 frames (1 s) [29], and only these frames are used for classification, without augmentation for the ones in between.

We train two types of classifiers. The first type is based on fully connected neural network architectures (artificial neural networks—ANN) that have as input one vector per sample. The second type is based on LSTMs to examine the temporal structure of the data, which we train on stacked sequences of vectors.

For both ANN and LSTM classifiers, we parametrize our experiments with respect to the object datasets. These are categorized based on:

- The **dataset combinations** for training and evaluating the classifier,
- The **object detector classes** and
- The **object detection thresholds**.

We categorize as such after considering our objective, i.e., to assess whether an object detector can be used as the first step in an indoor location recognition pipeline. In this context, we experiment with using either object annotations or detections to model the locations. At test time, we compare against object detections in order to compare the modeling capabilities offered from both train sets. Hence, the dataset combinations comprise the scenarios that affect the composition of a location classifier’s dataset. We consider *Labels to Detections* (L2D) which use the object annotations for training and the detections for testing and *Detections to Detections* (D2D) that contain only detections for both splits. For comparison, we also consider *Labels to Labels* (L2L) which consist of the object annotations for both splits; i.e., the object detections are *not* used.

The object detector variations were discussed in Sect. 6.3.2. Using this as a parameter means that we vary the object detector that produces the object dataset. As a result, different object classes are learned. This, in turn, leads to generating object vectors (BPV) of different lengths.



The detection threshold creates a trade-off between the confidence and the number of detections. Higher thresholds lead to more confident but fewer detections. Lower thresholds provide more objects, but with more false positives. In the D2D experiments, we always use detections from the same threshold for both training and testing.

### 6.3.4 Activities

For activity detection, we also rely on object detections [14]. We use the set of 20 object classes of ADL20 as described in Sect. 6.3.1 (Table 6.1), enhanced with object-related information. For every video frame, we extract objects along with their size and position in the frame. As features, we consider the BPVs as described in Sect. 6.3.3, along with the bounding box positions and the centralities.

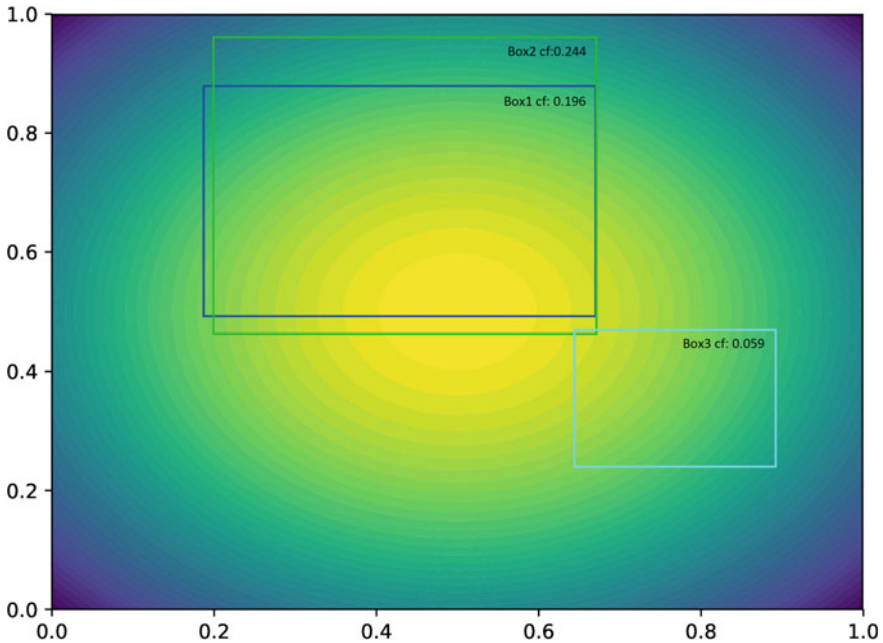
The bounding box (**BB**) position constitutes a 4-vector per object containing the ( $x$ ,  $y$ , width, height) parameters that characterize a bounding box. The values are normalized to the width and height of the frame to fall into the  $[0-1]$  range. For 20 object classes, the BB feature has length 80. The centrality feature (**CF**) signifies that a larger object area or a bounding box which is closer to the center of the image is more important for the detection of an activity. It constitutes a 2D Gaussian ( $\mu = 0.5$ ,  $\sigma = 0.1$ ) (in terms of normalized image coordinates) to produce a weight distribution that focuses its importance on bounding boxes found closer to the center of the frame. As a result, bigger boxes gain importance because they aggregate values over a larger area. Our intuition is that significant objects for human activities will be detected near the center of the scene or due to their size, they will draw attention to themselves [35]. A demonstration of the centrality feature's estimation is provided in Fig. 6.3.

## 6.4 Results

In this section, we delineate our experiments. Location classification is presented in Sect. 6.4.1 and activity recognition in Sect. 6.4.2.

### 6.4.1 Location Classification

We divide the experiments for location classification into ANN- and LSTM-based architectures in Sects. 6.4.1.1 and 6.4.1.3, respectively. In Sect. 6.4.1.2, we perform a per class examination for certain ANN cases.



**Fig. 6.3** Centrality value for three boxes. Box 1: 0.196, Box 2: 0.244 and Box 3: 0.059. As boxes become smaller or move away from the center, CF decreases

#### 6.4.1.1 ANN Classification

Our artificial neural network models consist of five fully connected layers, with rectified linear unit (ReLU) activations for the neurons of the input and three hidden layers. The neurons per layer are 64, 256, 128, 64 and 7 (for the output), respectively. We do not apply dropout, following preliminary tests where we experience slightly worse performance. We use categorical cross entropy to calculate the loss and stochastic gradient descent for optimization. All models are trained for 150 epochs. We set the starting learning rate at  $10^{-2}$  and divide by 10 every 50 epochs. Batch size is set to 64.

We implement experiments for the L2L, L2D and D2D cases of Sect. 6.3.3 with detection confidence threshold in the L2D and D2D cases ranging from 30 to 70%. The object sets vary between ADL20, ADL48 and MS COCO, with the latter only supporting the D2D case due to the lack of annotations for its object classes on the ADL dataset. The classifiers for each object set only differ in the input feature size which ranges between 20, 48 and 80, respectively. In Tables 6.6 and 6.7, we present the results in terms of overall Top1 accuracy and averaged F1-score over the seven locations in the test set, respectively.

When considering the *dataset combinations*, the highest classification accuracy is found in the L2L scenarios. This is expected since the object annotations do not

**Table 6.6** ANN Top1 accuracies (%)—averages of the best five models of each experiment. Comparisons between L2L, L2D and D2D for the various detector cases and detection thresholds. L2L outperforms the variants that depend on object detectors. Decreasing the object detection threshold improves classification accuracy for all object sets

| L2L     |         | Thresh. (%) | L2D     |         | D2D     |         |         |
|---------|---------|-------------|---------|---------|---------|---------|---------|
| ADL20   | ADL48   |             | ADL20   | ADL48   | ADL20   | ADL48   | COCO    |
| 77.7004 | 77.0448 | 30          | 59.6844 | 54.8024 | 62.9514 | 56.4688 | 64.3558 |
|         |         | 40          | 58.1134 | 48.1568 | 60.7436 | 55.8604 | 62.8764 |
|         |         | 50          | 56.6452 | 47.666  | 58.7338 | 55.1106 | 60.839  |
|         |         | 60          | 55.0006 | 39.368  | 55.7656 | 52.0564 | 57.8008 |
|         |         | 70          | 47.6106 | 38.953  | 51.6538 | 48.6618 | 51.7168 |

**Table 6.7** ANN F1-scores averaged over the seven location classes for the best performing model in Top1 accuracy. Comparison between L2L, L2D and D2D for the various detector cases and detection thresholds. The difference from the Top1 accuracies is attributed to the fact that certain locations (corridor and outdoor) are almost undetectable, affecting the average score

| L2L    |        | Thresh. (%) | L2D    |        | D2D    |        |        |
|--------|--------|-------------|--------|--------|--------|--------|--------|
| ADL20  | ADL48  |             | ADL20  | ADL48  | ADL20  | ADL48  | COCO   |
| 58.474 | 57.738 | 30          | 45.982 | 41.562 | 47.441 | 41.388 | 43.167 |
|        |        | 40          | 42.633 | 39.211 | 45.181 | 40.247 | 39.484 |
|        |        | 50          | 42.875 | 37.01  | 42.608 | 38.696 | 36.785 |
|        |        | 60          | 39.21  | 29.674 | 39.043 | 34.866 | 34.758 |
|        |        | 70          | 34.151 | 22.252 | 34.261 | 31.882 | 30.261 |

contain detector-induced noise, so the train set is clean with no objects out of place. When detectors are used, the D2D classifiers tend to outperform the L2D for the same object sets, even though they are trained on noisier samples. This fact provides insights about the way the ANN classifier handles noise. It will be confused by unexpected detections at test time; however, if it has faced similar samples during training, it deals with them more successfully at test time.

Varying the *object detector* affects the classification results significantly. In Tables 6.6 and 6.7, ADL20 L2D and D2D outperform their ADL48 L2D and D2D counterparts and COCO D2D performs better than both. When comparing ADL20 L2D with ADL48 L2D, it is important to consider the detection datasets. The test set for ‘ADL20’ consists of 67,906 ground truth boxes, and for ‘ADL48’ it is 95,845 (TP + FN in Table 6.8). The additional 28 k boxes of ‘ADL48’ belong to the harder to detect classes that are discarded from ‘ADL20’. The low average precision values for these classes (Table 6.5) indicate that most of their instances are not detected. This suggests a harder task for ‘ADL48’ to produce the ‘detections’ dataset for any confidence threshold. For example, at 50% confidence it has less TP but more FP and FN (Table 6.8). These can be interpreted as increased noise (FP) and reduced detection quality (FN) when compared to ‘ADL20’.

**Table 6.8** ADL20/48 object detector results. True positives decrease along with the false positives as the confidence threshold increases, complicating the classification task

| Detector<br>Thresh. (%) | ADL20  |        |        | ADL48  |        |        |
|-------------------------|--------|--------|--------|--------|--------|--------|
|                         | TP     | FP     | FN     | TP     | FP     | FN     |
| 30                      | 14,777 | 12,025 | 53,129 | 12,619 | 24,166 | 83,226 |
| 40                      | 13,277 | 8231   | 54,629 | 10,509 | 13,800 | 85,336 |
| 50                      | 11,762 | 5744   | 56,144 | 8493   | 8051   | 87,352 |
| 60                      | 9951   | 3784   | 57,955 | 6532   | 4558   | 89,313 |
| 70                      | 7621   | 2262   | 60,285 | 4417   | 2209   | 91,428 |

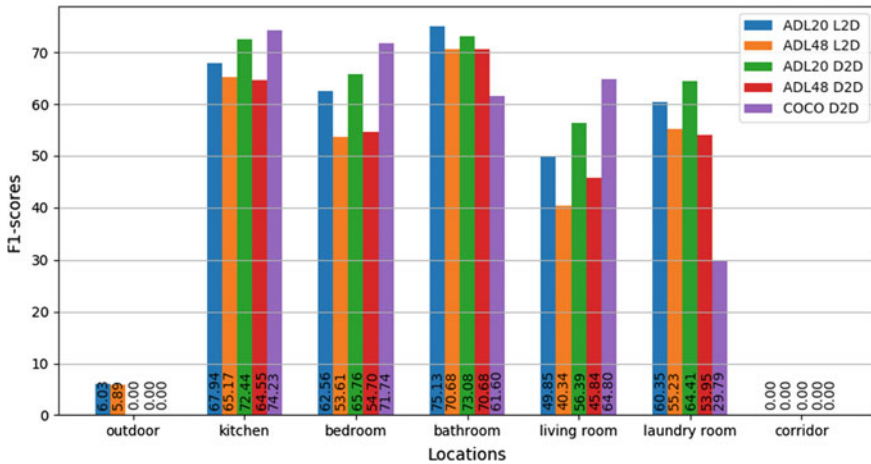
In terms of location classification accuracy, ADL20 L2D 50% is almost 9% better from ADL48 L2D 50% and ADL20 D2D 50% is 3.6% better from ADL48 D2D 50%. Finally, ‘COCO’, due to the higher number of training samples for each object class (over 5 k [44]) and despite consisting of 80 classes, is more robust in its object detections and the resulting location classifiers. Interestingly, in the L2L case the ADL48 variant is on a par with ADL20, meaning that the additional object classes, when not burdened by noise, do not harm location classification. The fact that COCO outperforms all other detector-based location classifiers adds to this, showing that quality detections without many false positives (resembling L2L as much as possible) even for classes from a more general context are useful.

We vary the *detection threshold* from 30 to 70% with a step of 10%. Our results suggest that as it increases, location classification performance drops. Lower thresholds lead to more available true positive object detections. This allows the location classifiers to identify uncertain locations easier, showing that they are resistant to noise. On the other hand, higher thresholds result in fewer detections with higher confidence on average and fewer false positives which, evidently, are not as adequate for inferring the location. The significant variance in the number and quality of detections as a result of modifying the confidence threshold for ‘ADL20’ and ‘ADL48’ is shown in Table 6.8 where we report the object detection results on the ADL test set videos.

#### 6.4.1.2 Examination Per Class

In Fig. 6.4, we compare the per class F1-scores for selected ANN classifiers to examine which locations are easier and which are harder to detect. No classifier is universally better, but superiority of certain classifiers can be observed for individual locations.

ADL20 outperforms ADL48 for all locations in both the L2D and the D2D cases. Similarly, the D2D cases outperform their L2D counterparts per class in most situations. COCO D2D performs best for ‘kitchen,’ ‘bedroom’ and ‘living room’ due to its ability to detect additional objects such as ‘fork,’ ‘sofa,’ ‘chair’ and ‘bed.’ However, it underperforms for ‘laundry room’ because it lacks a location-specific object class



**Fig. 6.4** Per class F1-scores at detection threshold 0.3 for five ANN classifiers (better seen in color—the order of names in the legend identifies their order in the graph)

related to the ‘washer/dryer’ of ADL20/48. Locations ‘outdoor’ and ‘corridor’ generally suffer due to the scarcity of training samples and explicitly associated objects (Table 6.2).

### 6.4.1.3 LSTM Classification

We are interested in studying the succession of objects in video segments instead of single frames. In Fig. 6.1 (left), the ‘kitchen’ scene lasts for 1260 consecutive frames (42 s) and the detected objects are not consistent throughout the segment. In certain views of the scene, the output of object detection is BPVs that cannot be associated with the ongoing location, for instance when no objects have been detected. Classifying one such BPV with an ANN classifier (e.g., ADL20 D2D 0.3) produces the mistaken prediction ‘laundry room,’ in between correct predictions of ‘kitchen’ for the surrounding frames. These frames include objects such as ‘fridge’ or ‘oven,’ but the frame in question does not. This observation drives our LSTM experiments in order to investigate how the temporal coherence of a scene can improve classification.

To test our hypothesis, we train an LSTM network with the dataset of ADL20 D2D 0.3. For training, we set the sequence size to 20 frames without augmenting the dataset with overlaps, so each sample is seen once per epoch as part of a single sequence. When testing the previously misclassified frames—now being part of a sequence—we find that the resulting location does not change from ‘kitchen’. Another interesting remark from this example is the ability of the LSTM to revert the prediction back to ‘kitchen’ if it happens to misclassify certain frames. Given a slice of three BPVs which contain only the ‘tap’ object and having from previous frames an ongoing location prediction of ‘kitchen’ with 52% probability, we classify the first BPV. It is classified

as ‘kitchen’, but its probability drops to 49%. The following ‘tap’-BPV modifies the prediction to ‘bathroom’ with probability 50%, and ‘kitchen’ drops further to 47%. This pattern continues for the third ‘tap’-BPV. However, given a vector that includes ‘fridge,’ the ‘kitchen’ prediction returns with increasing confidence, demonstrating the ability of the LSTM to recover from false intermittent predictions.

In order to test whether the LSTMs are also quantifiably better than ANNs, we repeat the dataset parametrization experiments from Sect. 6.3.3. We expect higher Top1 accuracies with LSTMs, as well as to confirm the relative associations between L2L, L2D, D2D, the object detectors and the detection thresholds.

For our experiments, we use two stacked LSTM layers and a fully connected layer, applied at the last sequence step of the second layer. We vary the feature size between 20, 48 and 80 following the BPV requirements. We set the number of hidden units to double the feature size. Following the ANN training scheme, we use categorical cross entropy to calculate the loss and stochastic gradient descent for optimization. All models are trained for 150 epochs with  $10^{-2}$  starting learning rate divided by 10 every 50 epochs. Sequence size is set to 20 which corresponds to a video duration of 20 s (i.e., 20 frames sampled at 1 fps) and batch size to 16 sequences.

At training time, we use a single label to describe a sequence. To produce it, we perform majority voting on the labels of all BPVs in the sequence and use that as the ground truth. Thus, the classifier is trained to produce a single prediction for the full sequence. At test time, we want to evaluate for every frame and not only once per sequence. To that end, we clone the prediction to the length of the tested sequence, to be able to evaluate against all the labels of the sequence one by one.

In Table 6.9, we report Top1 accuracies. For every task, the LSTM model surpasses the ANN equivalent. Except for the L2L combinations where the results are relatively close (2–4% difference), LSTMs show significant improvement, especially at the hardest cases, e.g., ADL48 L2D 0.6 (20.8%) and ADL48 L2D 0.7 (16.5%). The same conclusion can be drawn from Table 6.10 where the F1-scores are presented

**Table 6.9** LSTM Top1 accuracy (%)—averages of the best five models of each experiment. In parentheses, the differences from the respective ANN experiments

| L2L                |                    | Thresh.<br>(%) | L2D                 |                     | D2D                |                    |                     |
|--------------------|--------------------|----------------|---------------------|---------------------|--------------------|--------------------|---------------------|
| ADL20              | ADL48              |                | ADL20               | ADL48               | ADL20              | ADL48              | COCO                |
| 80.2384<br>(+2.54) | 80.6324<br>(+3.59) | 30             | 70.6992<br>(+11.01) | 63.8146<br>(+9.01)  | 70.1068<br>(+7.16) | 65.0716<br>(+8.6)  | 75.4906<br>(+11.13) |
|                    |                    | 40             | 66.8322<br>(+8.72)  | 63.8342<br>(+15.68) | 69.1038<br>(+8.36) | 62.655<br>(+6.79)  | 73.9324<br>(+11.06) |
|                    |                    | 50             | 68.8314<br>(+12.19) | 61.4836<br>(+13.82) | 67.1894<br>(+8.46) | 59.752<br>(+4.64)  | 73.1108<br>(+12.27) |
|                    |                    | 60             | 63.4306<br>(+8.43)  | 60.1648<br>(+20.8)  | 62.84<br>(+7.07)   | 59.3146<br>(+7.26) | 72.3696<br>(+14.57) |
|                    |                    | 70             | 61.7324<br>(+14.12) | 55.445<br>(+16.49)  | 61.8568<br>(+10.2) | 56.7314<br>(+8.03) | 67.0688<br>(+15.35) |

**Table 6.10** LSTM F1-scores averaged over the seven location classes for the best performing model in terms of Top1 accuracy. In parentheses, the differences from the respective ANN experiments

| L2L               |                   | Thresh.<br>(%) | L2D                |                    | D2D                |                    |                    |
|-------------------|-------------------|----------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| ADL20             | ADL48             |                | ADL20              | ADL48              | ADL20              | ADL48              | COCO               |
| 63.793<br>(+5.32) | 58.923<br>(+1.19) | 30             | 54.099<br>(+8.12)  | 54.154<br>(+12.6)  | 52.607<br>(+5.17)  | 51.421<br>(+10.03) | 53.543<br>(+10.38) |
|                   |                   | 40             | 52.259<br>(+9.67)  | 48.385<br>(+9.17)  | 53.782<br>(+8.6)   | 49.091<br>(+8.84)  | 51.45<br>(+11.97)  |
|                   |                   | 50             | 52.587<br>(+9.71)  | 47.171<br>(+4.16)  | 49.434<br>(+6.83)  | 46.28<br>(+7.58)   | 50.864<br>(+14.08) |
|                   |                   | 60             | 42.147<br>(+2.94)  | 45.754<br>(+16.08) | 46.914<br>(+7.87)  | 45.57<br>(+10.7)   | 50.065<br>(+15.31) |
|                   |                   | 70             | 46.012<br>(+11.86) | 39.938<br>(+17.69) | 47.002<br>(+12.74) | 41.196<br>(+9.31)  | 45.776<br>(+15.52) |

instead. As expected, the absolute values are lower, because they are influenced by the distribution of the dataset and affected by its imbalance.

## 6.4.2 Activity Classification

In Sect. 6.4.2.1, we present the results of the activity classification scheme and in Sect. 6.4.2.2 an analysis of the class confusions. All our tests consider the **detections to detections** dataset combination introduced in Sect. 6.3.3, where both train and test splits are built from the output of an object detector. This provides a more realistic scenario for smart-home application development, compared to the label-to-label combination which assumes ideal object detections. Despite recent improvements [46] in the state-of-the-art detectors, perfect detections are not yet feasible and flawless annotations in unseen environments require significant human labeling effort.

### 6.4.2.1 LSTM Classification

For the activity classification experiments, we train an LSTM network for the sequences of each feature combination of Sect. 6.3.4 targeting the 33 activity classes of the ADL dataset (Table 6.3). We prefer LSTM over artificial neural networks for their ability to incorporate temporal changes, compared to per frame classification schemes that do not consider objects seen in the past. We train a single-layer LSTM with 80 hidden units with a fully connected layer for the output. We set sequence size to 150 frames and batch size to 64. We apply 15% dropout with  $10^{-4}$  starting learning rate with polynomial decay down to  $10^{-6}$  in 1000 training iterations. We finish training after 1500 iterations.

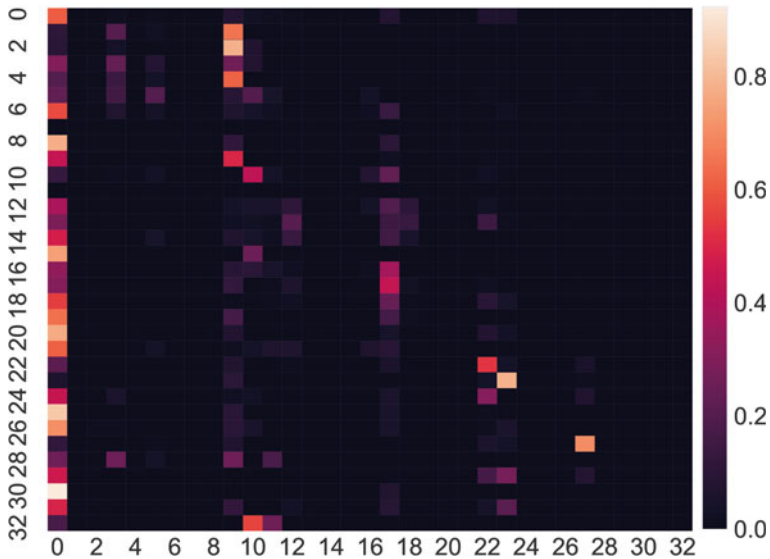
**Table 6.11** LSTM Top1 accuracies (%) and averaged F1-scores for all 33 classes for the feature combinations

| Feature | BPV   | BB    | CF    | BPV + BB | BPV + CF | BB + CF | BPV + BB + CF |
|---------|-------|-------|-------|----------|----------|---------|---------------|
| Top1    | 32.16 | 28.66 | 23.38 | 29.21    | 32.84    | 31.06   | <b>33.97</b>  |
| F1      | 10.51 | 9.07  | 6.69  | 11.11    | 10.89    | 10.83   | <b>12.4</b>   |

We report Top1 accuracies and F1-scores in Table 6.11. The highest performing individual feature is the binary presence vector. Adding the bounding box coordinates hurts results, but adding the centrality feature leads to the best performance overall. The high number of classes adds complexity to the classification task when compared to locations and leads to lower results overall.

#### 6.4.2.2 Class Confusions

The confusion matrix in Fig. 6.5 shows a specific trend. It suggests strong preference to certain activities, including 0: ‘background’, 3: ‘brushing teeth’, 9: ‘laundry’, 10: ‘washing dishes’, 12: ‘making tea’, 17: ‘making cold food/snack’, 22: ‘watching TV’, 23: ‘using computer’, 27: ‘reading book’. Beside their true positives, these classes attract false positives from conceptually relevant activities that rely on the same

**Fig. 6.5** Confusion matrix for BPV + BB + CF. Some activities can be assigned to semantic super sets



objects for recognition, but have fewer instances associated with them at training time (Table 6.3).

Class 17: ‘making cold food/snack’ contains false assignments from classes 16: ‘making hot food’ and 18: ‘eating food/snack.’ These activities rely on similar kitchen objects such as ‘dish’, ‘mug/cup’ and ‘tap’, but the classifier assigns them to the class with the most instances during training. Similarly, instances from 29: ‘writing’ are assigned to 27: ‘reading book’ based on ‘book’ as the detected object. Further confusions that regard semantic relevance include classes 1: ‘combing hair’, 2: ‘makeup’ and 4: ‘dental floss’ with 9: ‘laundry’, class 28: ‘mouth wash’ with 3: ‘brushing teeth’ and class 32: ‘grabbing tap water’ with 10: ‘washing dishes’.

## 6.5 Discussion

We envision a system that recognizes activities and locations from objects in a real setting. We structure the task in a very simple way, i.e., to solely rely upon the presence of objects in the scene for inference. This is a source of confusion even with the assumption of perfect detections, considering that objects are naturally found in multiple locations (‘door’) or are movable (‘cup’). We work with these limitations and explore ways to address them by relying on the temporal associations of objects to learn an improved representation of a scene or an activity.

Using the L2L combination for the location classifiers is not a pragmatic approach mostly due to the difficulties in data collection, such as human annotation effort in customized home environments. To mitigate these, we make use of automatic object detectors, pretrained on specific sets of object classes. Initially, this leads to the L2D case, where we can train on generic room representations; e.g., a common kitchen has a fridge, an oven and a tap, and expect to detect these objects in the test environment. However, the D2D classifiers have better performance than L2D and show increased resilience to noisy detections at test time. Additionally, they are more convenient installation-wise, since they abolish the necessity for labeling locations with objects. Thus, having minimized the required human labeling effort, it becomes easier to learn new representations of existing places (e.g., with a specialized detector that was previously unavailable), but also of unseen locations not included in the original categories.

Our purpose is to evaluate the applicability of the detectors in terms of object variability and acceptable accuracy of activity and location classification. The L2L experiments have the highest location classification performance and establish the idea that the amount of noise in the object detections (in this case, the lack thereof) influences the results substantially. A second significant outcome is that the variability of available objects can enhance the ability of a classifier to detect a location accurately. This is observed from ADL48 L2L which outperforms ADL20 L2L in Top1 accuracy in the LSTM tests. However, when the objects contain noise, less is more; i.e., in L2D and D2D, ADL48 does not outperform ADL20 in any (LSTM or ANN) experiment.

To enhance our original objective in activity classification, we consider a scenario with detections of house-related objects with additional object appearance features that are dynamic in time. We examine them with LSTM and find that this more complicated scenario cannot be sufficiently tackled with object-based features. While seven locations have been distinguished with up to 70% accuracy (LSTM ADL20 D2D 0.3), 33 activities reach 34%, with multiple intra-class similarities being observed and misclassified as such. The additional information about the object detections contributes to the results; however, it is not enough to properly address the elaborate task of analyzing semantically similar activities and to fully counter the bias toward classes with higher prior probability.

To our knowledge, there is no related work that tackles location classification in the ADL dataset. For the task of activity classification, related work does exist; however, the evaluation metrics and the regarded activity sets do not allow for a fair comparison. For example in [47], the authors report 26.3% average accuracy only for the 18 activities of the ADL dataset.

## 6.6 Conclusions

Throughout this chapter, we explore the recognition of indoor locations and human activities in egocentric videos. We utilize a state-of-the-art object detection architecture, trained separately on three object sets. We apply it on egocentric videos to extract objects at various detection thresholds and classify these detections with artificial neural networks and long short-term memory networks to infer locations or activities.

We find that the selection of object set affects the relevance of the detections in the location classification task and the detection threshold, their number and quality. Using the binary presence vector, we manage to have acceptable performance for indoor location classification, reaching 75.5%. One important discovery is that the lack of noise in the detections is preferable, but if it cannot be avoided the true positive/false negative trade-off favors the true positives even at the expense of extending the set of false positive detections. The comparison between ANN and LSTM promotes the incorporation of temporal structure in the BPVs (Tables 6.9 and 6.10) in order to capitalize on the sequential nature of the data and minimize the effects of erroneous detections.

We also find that the more complicated task of activity classification is harder to tackle based only on object-based features. Our results show that certain activities are easier to recognize than others, mostly due to their higher occurrence rate in the training set. We also find that activities which belong in semantic ‘super’ sets tend to be learned as belonging to the one representative activity that has the most instances in the training set.

An interesting direction for future work would be to analyze the associations between locations and activities and whether the first can assist in recognizing the latter.

**Acknowledgements** This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157.

## References

1. Kapidis G, Poppe RW, van Dam EA et al (2018) Where Am I? Comparing CNN and LSTM for location classification in egocentric videos. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), pp 878–883
2. Ma M, Fan H, Kitani KM (2016) Going deeper into first-person activity recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 1894–1903
3. del Molino AG, Tan C, Lim J, Tan A (2017) Summarization of egocentric videos: a comprehensive survey. *IEEE Trans Hum-Mach Syst* 47:65–76. <https://doi.org/10.1109/THMS.2016.2623480>
4. Yonetani R, Kitani KM, Sato Y (2016) Recognizing micro-actions and reactions from paired egocentric videos. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 2629–2638
5. Damen D, Leelasawassuk T, Mayol-Cuevas W (2016) You-Do, I-Learn: egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Comput Vis Image Underst* 149:98–112. <https://doi.org/10.1016/j.cviu.2016.02.016>
6. Kretch KS, Franchak JM, Adolph KE (2014) Crawling and walking infants see the world differently. *Child Dev* 85:1503–1518. <https://doi.org/10.1111/cdev.12206>
7. Nguyen T-H-C, Nebel J-C, Florez-Revuelta F (2016) Recognition of activities of daily living with egocentric vision: a review. *Sensors (Basel, Switzerland)* 16:72. <https://doi.org/10.3390/s16010072>
8. Karaman S, Benois-Pineau J, Megret R et al (2010) Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In: 2010 20th international conference on pattern recognition, pp 4113–4116
9. Teriús-Padrón JG, Kapidis G, Fallmann S et al (2018) Towards self-management of chronic diseases in smart homes: physical exercise monitoring for chronic obstruction pulmonary disease patients. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), pp 776–781
10. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05), vol 1, pp 886–893
11. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60:91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
12. Pirsivash H, Ramanan D (2012) Detecting activities of daily living in first-person camera views. In: 2012 IEEE conference on computer vision and pattern recognition, pp 2847–2854
13. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9:1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
14. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6517–6525
15. Furnari A, Farinella GM, Battiato S (2016) Temporal segmentation of egocentric videos to highlight personal locations of interest. In: Hua G, Jégou H (eds) *Computer vision—ECCV 2016 workshops*. Springer International Publishing, pp 474–489
16. Furnari A, Farinella GM, Battiato S (2017) Recognizing personal locations from egocentric videos. *IEEE Trans Human-Mach Syst* 47:6–18. <https://doi.org/10.1109/THMS.2016.2612002>
17. Nakamura K, Yeung S, Alahi A, Fei-Fei L (2017) Jointly learning energy expenditures and activities using egocentric multimodal signals. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6817–6826

18. Lee YJ, Ghosh J, Grauman K (2012) Discovering important people and objects for egocentric video summarization. In: 2012 IEEE conference on computer vision and pattern recognition, pp 1346–1353
19. Fathi A, Ren X, Rehg JM (2011) Learning to recognize objects in egocentric activities. CVPR 2011:3281–3288
20. Fathi A, Li Y, Rehg JM (2012) Learning to recognize daily actions using gaze. In: Fitzgibbon A, Lazebnik S, Perona P et al (eds) Computer vision—ECCV 2012. Springer, Berlin, pp 314–327
21. Poleg Y, Arora C, Peleg S (2014) Temporal segmentation of egocentric videos. In: 2014 IEEE conference on computer vision and pattern recognition, pp 2537–2544
22. Betancourt A, Díaz-Rodríguez N, Barakova E et al (2017) Unsupervised understanding of location and illumination changes in egocentric videos. *Pervasive Mob Comput* 40:414–429. <https://doi.org/10.1016/j.pmcj.2017.03.016>
23. Altwaijry H, Moghimi M, Belongie S (2014) Recognizing locations with Google Glass: a case study. In: IEEE winter conference on applications of computer vision, pp 167–174
24. Lee N, Kim C, Choi W et al (2017) Development of indoor localization system using a mobile data acquisition platform and BoW image matching. *KSCE J Civ Eng* 21:418–430. <https://doi.org/10.1007/s12205-016-1057-5>
25. Lu G, Yan Y, Sebe N, Kambhampettu C (2017) Indoor localization via multi-view images and videos. *Comput Vis Image Underst* 161:145–160. <https://doi.org/10.1016/j.cviu.2017.05.003>
26. Qian K, Zhao W, Ma Z et al (2018) Wearable-assisted localization and inspection guidance system using egocentric stereo cameras. *IEEE Sens J* 18:809–821. <https://doi.org/10.1109/JSEN.2017.2773487>
27. Dovgalecs V, Mégret R, Berthoumieu Y (2013) Multiple feature fusion based on co-training approach and time regularization for place classification in wearable video. *Adv Multimed* 2013. <https://doi.org/10.1155/2013/175064>
28. Vaca-Castano G, Das S, Sousa JP (2015) Improving egocentric vision of daily activities. In: 2015 IEEE international conference on image processing (ICIP), pp 2562–2566
29. Vaca-Castano G, Das S, Sousa JP et al (2017) Improved scene identification and object detection on egocentric vision of daily activities. *Comput Vis Image Underst* 156:92–103. <https://doi.org/10.1016/j.cviu.2016.10.016>
30. Greff K, Srivastava RK, Koutník J et al (2017) LSTM: a search space odyssey. *IEEE Trans Neural Netw Learn Syst* 28:2222–2232. <https://doi.org/10.1109/TNNLS.2016.2582924>
31. Karpathy A, Johnson J, Fei-Fei L (2015) Visualizing and understanding recurrent networks. arXiv preprint [arXiv:1506.02078](https://arxiv.org/abs/1506.02078)
32. Smith LN (2018) A disciplined approach to neural network hyper-parameters: part 1—learning rate, batch size, momentum, and weight decay. CoRR arXiv preprint [arXiv:abs/1803.09820](https://arxiv.org/abs/1803.09820)
33. Poppe R (2010) A survey on vision-based human action recognition. *Image Vis Comput* 28:976–990. <https://doi.org/10.1016/j.imavis.2009.11.014>
34. Bambach S, Lee S, Crandall DJ, Yu C (2015) Lending a hand: detecting hands and recognizing activities in complex egocentric interactions. In: 2015 IEEE international conference on computer vision (ICCV), pp 1949–1957
35. Bertasius G, Park HS, Yu SX, Shi J (2017) First person action-object detection with EgoNet. In: Proceedings of robotics: science and systems
36. Li Y, Zhefan Y, Rehg JM (2015) Delving into egocentric actions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 287–295
37. Poleg Y, Ephrat A, Peleg S, Arora C (2016) Compact CNN for indexing egocentric videos. In: 2016 IEEE winter conference on applications of computer vision (WACV), pp 1–9
38. Fathi A, Farhadi A, Rehg JM (2011) Understanding egocentric activities. In: 2011 international conference on computer vision, pp 407–414
39. Wray M, Moltisanti D, Mayol-Cuevas W, Damen D (2016) SEMBED: semantic embedding of egocentric action videos. In: Hua G, Jégou H (eds) Computer vision—ECCV 2016 workshops. Springer International Publishing, pp 532–545
40. Wu J, Osuntogun A, Choudhury T et al (2007) A scalable approach to activity recognition based on object use. In: 2007 IEEE 11th international conference on computer vision, pp 1–8

41. Su Y-C, Grauman K (2016) Leaving some stones unturned: dynamic feature prioritization for activity detection in streaming video. In: Leibe B, Matas J, Sebe N, Welling M (eds) *Computer vision—ECCV 2016*. Springer International Publishing, pp 783–800
42. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, pp 779–788
43. Deng J, Dong W, Socher R et al (2009) ImageNet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*, pp 248–255
44. Lin T-Y, Maire M, Belongie S et al (2014) Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) *Computer vision—ECCV 2014*. Springer International Publishing, pp 740–755
45. Girshick R (2015) Fast R-CNN. In: *2015 IEEE international conference on computer vision (ICCV)*, pp 1440–1448
46. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. *CoRR arXiv preprint [arXiv:abs/1804.02767](https://arxiv.org/abs/1804.02767)*
47. Nguyen T-H-C, Nebel J-C, Florez-Revuelta F (2018) Recognition of activities of daily living from egocentric videos using hands detected by a deep convolutional network. In: Campilho A, Karray F, ter Haar Romeny B (eds) *Image analysis and recognition*. Springer International Publishing, pp 390–398

# Chapter 7

## Improving the Collection and Understanding the Quality of Datasets for the Aim of Human Activity Recognition



Angelica Poli, Susanna Spinsante, Chris Nugent and Ian Cleland

**Abstract** In the last few decades, life expectancy has been increasing. This has resulted in a higher proportion of older adults and increased prevalence of chronic conditions, posing challenges facing care needs. A possible solution is to foster both the prevention and health-related re-education, supporting healthier lifestyle and facilitating independent living. To facilitate this, it is crucial to measure individual's key health metrics. For instance, human activity recognition through sensors provides valuable information about an individual's lifestyle. Some crucial decisions, among which the quality of data collection, strengthen the methodological approach. This chapter addresses how the quality of data may affect the recognition performance. Two datasets of daily activities were collected through a triaxial accelerometer placed on the subject's dominant wrist. The first dataset was collected by 141 users, whereas the second one comprised semi-realistic activities executed by three individuals. Specifically, outcomes were based on a comparison of activity recognition performance of six machine learning classifiers. Results show that, firstly, a higher number of features may not improve the recognition rate. Secondly, one approach may be robust in a laboratory setting but not generalizable to real-world applications. Finally, a great variability may increase the generalization of classifiers for successful activity recognition.

**Keywords** Activity recognition · Dataset quality · Features selection · Accelerometry

---

A. Poli (✉) · S. Spinsante  
Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, Italy  
e-mail: [a.poli@staff.univpm.it](mailto:a.poli@staff.univpm.it)

S. Spinsante  
e-mail: [s.spinsante@staff.univpm.it](mailto:s.spinsante@staff.univpm.it)

C. Nugent · I. Cleland  
School of Computing, Ulster University, Belfast, UK  
e-mail: [cd.nugent@ulster.ac.uk](mailto:cd.nugent@ulster.ac.uk)

I. Cleland  
e-mail: [i.cleland@ulster.ac.uk](mailto:i.cleland@ulster.ac.uk)

## 7.1 Introduction

Over the last few decades, life expectancy has been increasing. According to the World Health Organization (WHO, 2018), the number of individuals over 60 years is expected to double by 2050. Many challenges have arisen in response to population ageing-related issues, among which the increase of chronic diseases and care needs are included. An impressive 58% of WHO Member States report having an e-Health support strategy or policy (WHO, 2016). Indeed, the integration of ICT and electronic systems known as e-Health plays an important role responding both to the need of rationalizing the healthcare costs and the aim of increasing the quality of healthcare processes. The relevance is evident in a wide range of healthcare applications, including many tools for health monitoring, disease prevention, diagnosis, treatment and lifestyle management [2, 31]. All of these scenarios are possible and efficient because ICT encompasses an extensive domain of technology applications that facilitate the capturing, processing, storage and exchange of data via electronic communication [34]. Additionally, in order to support a safer and healthier lifestyle, the development of smart home-based health care applications can create an intelligent ambient for comprehensive safety monitoring, recognition and tracking of human activities at home.

The domain pertaining to recognizing and classifying patterns of the human activities is human activity recognition (HAR). HAR is the process whereby the behavior of an individual and their own environment are monitored and analyzed to infer the underlying user activities [7]. The typical sequence of processing steps in HAR (i.e., data acquisition, data filtering, data segmentation, features extraction and classification) is one of the most commonly used workflows for developing automated activity recognition models. Such automated recognition models of activities can be used to understand the relationship between the behavior and the health status of an individual. For example, many behavioral changes (e.g., the manner and speed to perform activities) in older people may be correlated with the onset of a specific disease [10]. As a consequence, some recent works emphasize the capability of HAR in smart environments as a valuable assistant aimed at health monitoring, anticipating and avoiding undesirable consequences [20]. This integration can be an important step toward monitoring the health of individuals living longer and independently at home.

Some recent works have underlined significant progresses on this subject. However, the most discussed limitation is the nature and quality of the collected data. This issue arises from the assumption that many existing studies are conducted in a laboratory setting, under highly specific instructions, including simulated activities [33]. As a consequence, data collected in the laboratory setting may be not totally representative of a real-world scenario. Indeed, within a group of people, each individual performs different activities in a different manner, but also to him/herself over time [6, 33]. As a result, one approach may be effective in a laboratory setting, but may face some limitations in real-world applications [17].

Therefore, based on a previous work about a HAR solution, exploiting a large-scale dataset [9], this chapter investigates how the data quality may affect the

classification results, by resorting to the use of data collected both inside and outside the laboratory. The nature of this new approach aims to provide a more reliable analysis, thanks to the more realistic dataset collected outside the laboratory. Specifically, the main contribution of this work is to assess and compare the recognition performance both by training and by testing a HAR classifier on simulated data and on data collected in a more realistic setting.

The chapter is organized as follows: Sect. 7.2 clarifies the motivations of this work. Section 7.3 presents the background about the role that Information and Communication Technology (ICT), HAR processes and smart sensors may have in supporting the health system and their possible impact. Section 7.4 explains the data acquisition modality and the dataset structure. Section 7.5 details the system design about the pre-processing, including data labeling and segmentation, filtering and outlier's analysis. Section 7.6 shows three case studies and the corresponding results. Finally, Sect. 7.7 draws the conclusion and highlights several developments for future research directions.

## 7.2 Motivation

The inspiration for the research presented in this chapter lies in a work, where the HAR was assessed in laboratory setting under specific instructions. In particular, the individuals involved in the experiments considered here performed the same set of common daily activities analyzed by Cleland et al. [9].

The overall aim of the activity was to study how the nature and quality of data collected may affect the classification results [26]. For this purpose, we collected a dataset of human daily activities in a real-world setting, ensuring a more realistic dataset than in a lab scenario. In this context, we focused on observing the performance of several machine learning approaches by comparing the two datasets, collected under different conditions. More specifically, the accuracy of six among the most commonly used classification methods was assessed using the Weka Explorer toolbox (the University of Waikato, version 3.8.1). Weka currently includes many different machine learning schemes but in this specific work, the Decision Tree (implemented as J48 in Weka), Random Forest, Support Vector Machine, Naïve Bayes, Neural Networks and k-Nearest Neighbor systems were implemented and assessed. In particular, they are compared in terms of correct classification rate, F-measure and confusion matrix, to evaluate their performance in both the analyzed datasets.

The two datasets, named Dataset 1 and Dataset 2, referred to laboratory and realistic setting, respectively, provided crucial results for the following objectives:

- To extend the features subset, aiming to capture more information about the activities;
- To perform a cross-test validation, with the training phase on Dataset 1 and the testing phase on Dataset 2, in order to examine the algorithm functioning;



- To test the learning algorithms over both the datasets combined to obtain a general classifier.

### 7.3 Background

The development of ambient intelligence-based health care applications is the current approach for alleviating several problems healthcare systems are facing. In particular, ICT has become vital to manage the increasing costs and also improve the quality of health care. Two examples are the reduction of medical errors thanks to smart systems [5] and the monitoring of patients at home with technological solutions enabling safety recognition and tracking of human activities [29]. Among the others, HAR is complex process that aims to recognize human behaviors starting from the monitoring and analysis of data acquired by smart sensors [7, 16].

Sensors for activity monitoring have been studied since 1999 [25]. However recently, the developments in both technology and care models are extending the adoption of sensors by researchers who want to address HAR, especially with wearable sensors (e.g., accelerometers, magnetometers, gyroscopes, etc.) that have minimal invasive effects on daily life [4, 21, 28]. According to Khusainov et al. [15], on-body sensing is the most common monitoring technology for gait assessment, fall detection and activity recognition and classification. The most broadly used wearable sensor for the purpose of HAR is the accelerometer [4, 35]. It is unsurprising that triaxial accelerometer sensors have been increasingly popular, being a user-friendly (i.e., small size and light weight), robust, low-cost and highly reliable device, that can also provide high accuracy [23].

Particular attention needs to be paid regarding the selection of sensor placement, as the human body activity is a coordinated movement of several body parts and properly connected joints. For example, in the study by Cleland et al. [9], the individuals wore the sensor on the dominant wrist, instead Gupta and Dallas [12] placed the accelerometer on the waist. While in this latter work, the most prevalent daily activities (sitting, standing and walking) have been well-recognized, and in the former one, the similarity of the arm movements caused confusion among different activities (e.g., walking and stepping). As one single accelerometer may be not sufficient to differentiate and distinguish all the investigated activities, some studies rely on a larger number of sensors [27]. Contrarily, other studies achieved high classification accuracies at 95.0% with one ankle sensor [22]. Indeed, Cleland et al. [8] show that the difference in accuracy of activity detection, with using one or multiple accelerometers, is not as high as expected, especially considering activities grossly different from each other in terms of movements.

After selecting the optimal number and position of sensors, a specific sequence of processing steps is needed to detect and recognize activities. The typical workflow (Fig. 7.1) of HAR is known as Activity Recognition Chain (ARC), as accurately detailed in [3].



**Fig. 7.1** Typical Activity Recognition Chain (ARC) to recognize activities from wearable sensors

Each phase in the ARC is crucial. However, the development of an optimal HAR system starts collecting an adequate human activity database. According to Vrigkas et al. [33], the quality of the input data is one of the most important aspects to consider. For this reason, in this chapter we discuss whether and how both the feature extraction and the nature of data acquired may influence the activity recognition and classification. In particular, this latter issue arises from the assumption that there is no a common definition or structure of human activities to support a common problem statement, because the way humans perform the same activity is highly diverse. According to Bulling et al. [6] and Kong and Fu [17], the intraclass and intraperson variability in performing a specific activity are among the common research challenges. Most past works, for instance [9, 11, 19], have collected data by instructing subjects to perform activities, thus missing the great variability. The scripted activities result in simulated and un-natural behaviors with small variability among participants. On the contrary, Vaizman et al. [30] proposes a context recognition *in-the-wild* condition, capturing the user's natural behaviors instead of simulated ones. This means that, promoting real-life applications, users were in an uncontrolled environment using the device in a spontaneous fashion, without forcing the tasks or instructions. Behavioral models that fit well to recognize simulated activities may poorly generalize *in-the-wild*, failing in such applications.

In the last phase (Classification) of ARC, the development and analysis of the classification algorithms are important tasks. The main idea is to create a machine that learns the system automatically, using the designed algorithms that iteratively learn from data [1] and generate behavioral models, without any human interventions.

## 7.4 Experimental Methodology

### 7.4.1 Data Acquisition and Datasets Description

The following section details the plan and the design methodology for the current study.

The most relevant reference highlighting the objectives of this activity is the previous work by Cleland et al. [9]. Their study involved 141 students for creating an

activity dataset (named Dataset 1). Within this, each student was assigned to one out of six scenarios (Table 7.1) selected to represent common activities of daily living (ADL).

Students were provided with video instruction describing each step of sensor calibration, positioning and data collection, including the manner in which to perform the activities. Nevertheless, students were not supervised during any of these processes [9].

The study herein described involved three additional individuals for collecting a second database (named Dataset 2). In order to obtain a comparable amount of data with Dataset 1, each individual was equally assigned to all six scenarios.

In both the datasets, information about the activity performed was obtained from a single accelerometer (Shimmer 2R, Shimmer Research, Dublin, Ireland) placed on the subject's dominant wrist. Data were recorded for two minutes per activity via Bluetooth. In the study recently performed, some challenges were applied to evaluate how the different data collection and data analysis may influence the HAR performance. For this reason, the three involved individuals were provided with prior calibrated device, instructions on how correctly wear the accelerometer and with real tools (e.g., an iron and a ball) to perform the different activities. Although no instructions on how to perform the activities were given, the individuals were supervised during the data collection, in order to annotate the data including any anomalies (i.e., the accelerometer breaks away from the wrist and falls on the ground) or unpredictable user's behavior. These challenges aimed to maintain the entirety of individuals' real attitude.

Note that activity recognition in real-world settings is much more challenging with respect to laboratory settings, and it may include also more confused activity events that need to be evaluated.

**Table 7.1** Scenarios, activities and corresponding number of assigned participants

| Scenarios          | Activities                                       | No. part. (dataset 1) | No. part. (dataset 2) |
|--------------------|--|-----------------------|-----------------------|
| Self-care          | Hair grooming, washing hands, brushing teeth     | 24                    | 3                     |
| Exercise (cardio)  | Walking, jogging, stepping up                    | 23                    | 3                     |
| House cleaning     | Ironing clothes, washing windows, washing dishes | 25                    | 3                     |
| Exercise (weights) | Arm curls, dead lift, lateral arm raise          | 21                    | 3                     |
| Sport              | Bounce ball, catch ball, pass ball               | 25                    | 3                     |
| Food preparation   | Mixing food, chopping vegetables, sieving flour  | 23                    | 3                     |

*No. part.* number of participants

### 7.4.2 Processing Steps

The success of any sensor application and data analysis solution depends on the data quality. Without quality, the value of data significantly decreases, and the diagnostic value is limited. Therefore, the data quality assessment is an important and integral part of any sensor application process, starting from the sensor system design to the data processing. For instance, body movements and user negligence in performing activities critically increase the probability of sensors' displacement or fall, causing interferences and calibration issues during the data acquisition.

Because of these listed issues, it is very difficult to analyze and recognize activities directly from collected data, without using any data pre-processing techniques.

As our first step, the name of activity labels was checked for unifying the individual's files. Each user timed the performed activity, using an own stopwatch for 2 min. Accelerometer data, collected between the start and the stop times, were saved in a CSV file with a matrix structure. Each user, after the data collection, independently labeled the file with the name of the scenario and the corresponding activity. Hence, a label was manually added and associated with the activity saved in the file.

Generally, the data preparation includes also the pre-processing phase, manipulating data to be suitable for extracting features. In theory, data emerging from this process contains less incomplete, noisy or inconsistent information. In this specific work, to reduce the noise at high frequency and improve the data quality, the accelerometer signal was filtered using both the median and low-pass Butterworth filter (4th order) with the cutoff frequency set at 15 Hz. The choice of this specific cutoff frequency arises from the assumption that 99% of the frequencies observed in human body motion are below 15 Hz [1].

Human motion has a fluid nature, and in such a way one activity blurs into the consecutive one [10], generating a continuous stream of information. A data segmentation procedure allows to split such time series into segments as discrete meaningful units of activity to improve the activity detection. Here, according to the previous study, the acceleration signals were split into non-overlapping windows with a fixed size of 4 s, resulting in 30 windows for each file. The selected window size was a compromise between small and large size, and it provided enough representative data in each window [22]. However, since the involved individuals were standing while pushing the start and stop times, the first and last windows were discarded. Such small recording errors produced discrepancies in the data size and/or calibration. These two issues are generally examined in the outlier analysis. In this case, the outlier detection consisted in a visual inspection method to identify anomalous data. The calibration issue, associated with an improper calibration or device malfunction, was evident when the sensitivity of the device was out of the maximum device range ( $\pm 1.5 \text{ g m/s}^2$  or  $14.71 \text{ m/s}^2$ ). As a result, the files exhibiting this problem were separately studied. To investigate the impact of the collected data quality in the development of the HAR models, we compared the accuracy of the classifiers using three different sets of features extracted from the accelerometer signals, as detailed in Table 7.2. Set 1 contained features selected from Cleland et al. [9]. Set 2 included

**Table 7.2** List of features used in this study

| Feature set | Time domain   | Frequency domain                      | Total no. features |
|-------------|---|---------------------------------------|--------------------|
| 1           | Mean, standard deviation, maximum, minimum, range, root mean square, signal magnitude area                            | Energy, entropy                       | 9                  |
| 2           | Set 1 combined with the following features: Correlation axis [24], Skewness [36], Kurtosis [36], Autocorrelation [10] | Location peaks [32], width peaks [32] | 15                 |
| 3           | Best features from Set 2 selected using Relief-F subset evaluation  |                                       | 11                 |

features tested in previous works about the HAR primarily in both time and frequency domain in accelerometer-based activity recognition. Set 3 combined features from Set 1 and Set 2 where 11 features that had the highest accuracies from each scenario were selected. Specifically, within the time domain, the information was computed individually for each of the three ( $x$ ,  $y$  and  $z$ ) components of the triaxial accelerometer signal. However, to minimize the effects of sensor orientation, the signal magnitude vector (SMV) was also considered, being less sensitive to orientation changes of the sensor node. SMV is extracted by combining the three axes, as defined by Eq. (7.1).

$$SMV = \sqrt{a_x^2 + a_y^2 + a_z^2} \quad (7.1)$$

On the contrary, within the frequency domain, the spectral features were calculated from the fast Fourier transform (FFT) of the SMV.

All defined subsets included a mixed set of both time and frequency features that is the effective technique to obtain high levels of classification accuracy, as suggested in some previous studies, for instance [4, 27].

## 7.5 Experiments and Performances Evaluation

The following section illustrates the three performed experiments and explains the corresponding results based on the evaluation of activity recognition performance of six classifiers, chosen according to the previous work [9]. These results demonstrate that some decisions are crucial for successful recognition, especially the quality of data collection that may affect the classifier's performance.

### 7.5.1 Extension of Features Subset

The design of features subset in a HAR system is a critical process, especially in a dataset with a great variability [6]. One may expect that more extracted features should result in a more discriminating power by the recognition algorithm. However, selected features in typology and number may affect the performance of a recognition system [12]. To analyze the impact of the performed choices, three different subsets of features were extracted from both the considered datasets and the performance of six machine learning classifiers was evaluated, applying a 10-fold cross validation partitioning.

Once the type of features is established, the issue becomes how to select the number of features that may affect both the time for developing the model and the recognition performance. Generally, a research strategy is used to find an optimal number of extracted features that fit the different classifiers and the specific applications.

For this reason, the potential optimal subset of features was examined in depth for correctly discriminating the human activities involved in this specific study. All the results obtained using the several classifiers are detailed in Tables 7.3 and 7.4, for both the Dataset 1 and Dataset 2, respectively.

First of all, the algorithm originally developed for the Dataset 1 was tested on the collected Dataset 2. This means that the same 9 extracted features (Set 1) of the previous work [9] were considered for the Dataset 2. This choice arose to understand how the selected classifiers change their performance in the presence of a pretended activity-based dataset, or a semi-realistic dataset, respectively.

**Table 7.3** List of classifiers and corresponding accuracy (%) using the three different features subset in Dataset 1

| Classifiers            | Set 1 | Set 2 | Set 3 |
|------------------------|-------|-------|-------|
| Decision Tree          | 85.51 | 56.88 | 56.96 |
| Random Forest          | 93.63 | 67.65 | 67.44 |
| Support Vector Machine | 77.91 | 52.85 | 52.18 |
| Naïve Bayes            | 60.92 | 38.35 | 37.38 |
| Neural Networks        | 82.35 | 57.52 | 57.57 |
| k-Nearest Neighbor     | 92.36 | 61.49 | 63.57 |

**Table 7.4** List of classifiers and corresponding accuracy (%) using the three different features subset in Dataset 2

| Classifiers            | Set 1 | Set 2 | Set 3 |
|------------------------|-------|-------|-------|
| Decision Tree          | 88.17 | 72.36 | 73.02 |
| Random Forest          | 93.82 | 80.34 | 80.64 |
| Support Vector Machine | 72.38 | 64.32 | 63.61 |
| Naïve Bayes            | 73.23 | 51.64 | 55.63 |
| Neural Networks        | 86.02 | 68.91 | 69.48 |
| k-Nearest Neighbor     | 88.32 | 64.72 | 70.10 |

The results obtained by training and testing on Dataset 1 and Dataset 2, respectively, using the original 9 features were promising, as it can be noticed in Tables 7.3 and 7.4. An average accuracy of 82% was achieved concerning the Dataset 1, while an average accuracy of 83% concerning the more realistic dataset (Dataset 2), especially for Random Forest and k-Nearest Neighbor. Comparing the results for each classifier from both the datasets, there was a surprising discrepancy: Four out of six classifiers worked better in Dataset 2 than in Dataset 1. A possible motivation of these unexpected higher percentages may be a consequence of the small data size of Dataset 2 with respect to the original one. These results will be compared with the results of the second experiment to analyze how the number of considered features may affect the achievement of classifiers.

To improve the recognition rate of an activity recognition model, we extended the original feature subset, resulting in 15 features (Set 2). Specifically, six features are added in both time and frequency domain, namely correlation axis, skewness, kurtosis, autocorrelation, location peaks and width peaks. Using the new subset of features, it is interesting to notice how the accuracy of each classifier is drastically decreased by a minimum of 8% up to a maximum of 28%, as shown in Tables 7.3 and 7.4.

Although some features may be very discriminative in conjunction with other features, they could generate low scores. For instance, missing to take into account features dependency may affect the performance of a feature selection algorithm. As a result, such algorithm assigns improper and similar rank to several features, causing redundancy. Looking at the theory and the definition of the six used machine learning algorithms, irrelevant and redundant training information may adversely influence their recognition performance.

For this reason, we shifted our focus on another effort: a number of features between the original one and the extended set, namely 11 features (Set 3). To evaluate the influence of the feature subset on the classifiers, three different feature selection algorithms were tested in Weka, namely the Correlation-based Feature Selection [13], Information Gain [14] and Relief-F [18] techniques. The performance of these techniques was estimated in three steps. Firstly, the rank of all fifteen features was computed. Secondly, the best-ranked features for each environmental context (laboratory and semi-naturalistic) were heuristically selected both in time and frequency domain. Thirdly, according to both Hall [13] and John et al. [14], the Decision Tree classifier was fed with such features to compare the obtained accuracy percentages with the feature sets. After evaluating the performance of these three feature selection methods in both datasets, the Relief-F method was found to be the most powerful algorithm for this purpose.

According to the above-mentioned outcomes, after implementing the Relief-F algorithm, the eleven features with the highest rank were chosen from both the two datasets. Unfortunately, two important disadvantages are associated with the Relief-F algorithm, namely the expensive computational cost and the potential failure in removing the redundant features. As a consequence, many approaches did not highly improve their own accuracy level.

Among the three efforts at improving the achievements of classifiers by testing different set of features, the best performance for both the two datasets was obtained using the Set 1. Consequently, the next two experiments were performed using the same type and number of features selected in the previous work by Cleland et al. [9].

### 7.5.2 Cross-Test: Training on Dataset 1 and Testing on Dataset 2

Recent studies are looking for many progresses in processing strategies to recognize basic daily human activities. However, they often conduct controlled experiments that miss the great variability of natural human behavior.

Getting out of the lab, Vaizman et al. [30] defines the human behavior by the wording *in-the-wild*. *In-the-wild* means in real-life settings in which people spontaneously conduct their daily behavior. The realistic behavioral context is extremely wide, complex and unpredictable: people walk, work or interact in different manners and typically focus on more than one activity at the same time [10]. This means that, when analyzing both data and model implementation approaches, it is also important to generate a model and validate it *in-the-wild*, making it appropriate for working in a real context.

For this reason, in the second study we focused on a cross-test. This experiment implied the training phase on the original dataset (Dataset 1) and the testing phase supplying the Dataset 2 as test set. In other words, the classification algorithms were trained with pretended activities and then validated in the semi-realistic dataset, with the majority of data not controlled and variable. A not surprising discrepancy between the performance of Dataset 2 (semi-realistic dataset) and of Dataset 1 (laboratory setting) was observed, as reported in Table 7.5.

By testing in daily life, the accuracy of all models dropped by a minimum of 10% (Naïve Bayes) up to 40% (k-Nearest Neighbor and Decision Tree). Basically, the

**Table 7.5** List of classifiers and corresponding accuracy (%) by training and testing on Dataset 1 and by training on Dataset 1 and testing on Dataset 2

| Classifiers            | Accuracy                                  | Accuracy                                  |
|------------------------|---|---|
|                        | Testing: Dataset 1<br>Training: Dataset 1 | Testing: Dataset 1<br>Training: Dataset 2 |
| Decision Tree          | 85.51                                     | 47.88                                     |
| Random Forest          | 93.63                                     | 59.78                                     |
| Support Vector Machine | 77.91                                     | 55.71                                     |
| Naïve Bayes            | 60.92                                     | 47.95                                     |
| Neural Networks        | 82.35                                     | 55.85                                     |
| k-Nearest Neighbor     | 92.36                                     | 52.05                                     |



algorithm, that had previously learned some regular activities in Dataset 1, resulted confused when it had to recognize activities differently performed in Dataset 2. On the contrary, as evident when training and evaluating the algorithms on the same dataset, they obtained an average accuracy greater than 80%. Further confirmation was observed from precision and recall (sensitivity), specifically from their harmonic mean (F-measure), that was computed for each classifier. The trend of F-measure values followed and confirmed the accuracy trend, achieving simultaneously the highest accuracy and F-measure with Random Forest approach.

Furthermore, the F-measure was calculated also for each single activity in the six different cases, as detailed in Fig. 7.2.

It is interesting to note how the highest values were achieved by the Lateral Arm Raise and Arm Curls activities. These two activities are part of the Exercise (Weight) scenario that is a workout. Indeed, this scenario included three mostly constant physical exercises that are performed almost in the same manner by everyone. Contrarily, the lowest values were reached by Mixing Food and Chopping Vegetables activities. These two activities are part of Food Preparation Scenario, and they represent almost entirely random activities. Mixing Food and Chopping Vegetables activities together with Washing Dishes and Washing Windows are usually performed in a different manner by different subjects, with a great variability. The skewed distribution of classes in daily life, which are predominantly natural activities, may affect the F-measure that is very sensitive to class skew.

Another point of interest is the Bounce Ball activity, for which the F-measure was always zero. The possible motivation is that the recognition of the Bounce Ball activity fails because in collecting the Dataset 2 a real ball was used by the volunteer, during the activity execution. The interaction with real objects (like the ball) instead

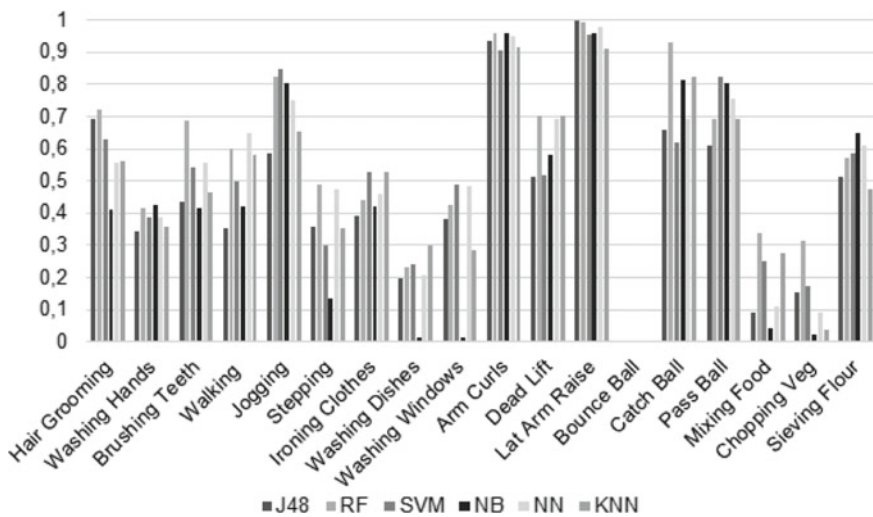


Fig. 7.2 F-measure values computed for each activity using the six classifiers

of pretending to use them may largely influence the intensity of arm movements in real-life settings, requiring more force for performing the activity. As a result, the strength of the acceleration signals highly increases with respect to the pretended activities (i.e., Dataset 1), and the classifiers may not recognize the previously learned activities.

These considerations were also analyzed from the confusion matrix point of view (Fig. 7.3).

Generally, the confusion matrix is a graphical tool that summarizes how the activities are classified. The rows represent data with their own activity label, and the columns represent the data with the activity label assigned by the classifier. The diagonal indicates the correct classifications.

Examining the performance at classifier level, the confusion matrix was analyzed for the Random Forest classifier due to its highest recognition rate. Figure 7.3 suggests that some specific misclassifications were more frequent than others:

- Activities of Walking (d), Lateral Arm Raise (l) and Pass Ball (o) represent similar, rhythmic and repetitive movements and were the best recognized ones;
- Activities with more subjective movements and gestures that change from one person to another are not well-recognized and often they are confused with other activities (e.g., Stepping (f) classified as Walking (d), Chopping Vegetables (q) classified as Ironing Clothes (g), Washing Hands (b) classified as Sieving Flour (r) and Jogging (e) classified as Pass Ball (o));
- Activities of Washing Hands (b), Washing Dishes (h), Hair Grooming (a), Chopping Vegetables (q) and Mixing Food (p) are the most challenging activities due to the great degree of random variability in performing them.

This clear difference in recognizing the daily life activities and laboratory activities remarks that laboratory data has a limited capability of representing the activities

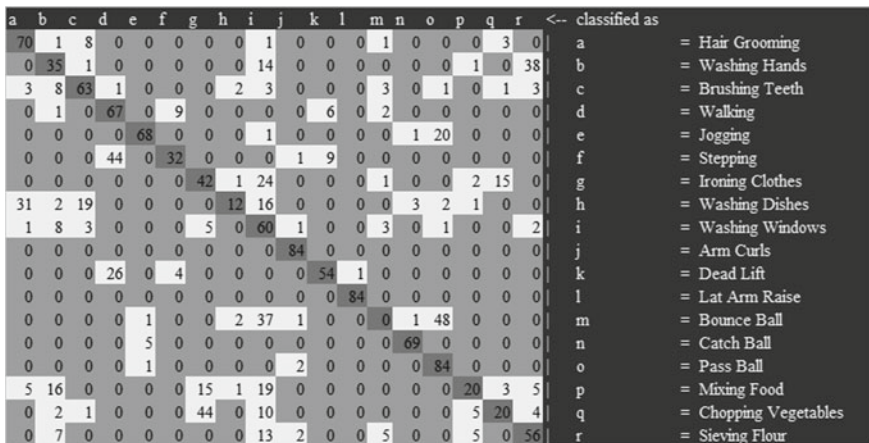


Fig. 7.3 Confusion matrix computed for each activity using only Random Forest classifier

undertaken in real daily-life settings. One approach may be effective in a laboratory setting but may have many difficulties in real-world application scenarios.

### 7.5.3 Combination of Datasets

The last experiment investigates the possibility to obtain a general classifier that works equally well on both datasets.

A third dataset was created for the last experiment. Dataset 3 was the combination of Dataset 1 and Dataset 2. Particular attention needed to be paid to divide the dataset in training and testing sets. Splitting randomly the dataset, the algorithm may be trained and tested on the same data. This problem may create misunderstandings, resulting in apparently higher accuracy. Therefore, in this specific work, the ratio of dataset splitting was set at 80–20%. This means that 80% of the dataset was used as training set and the remaining 20% as validating set, especially excluding the training subjects from the testing set.

Results show that the recognition rate of each classifier may improve by using data from both collections. As expected, in some cases, the accuracy was lower than the results obtained in the first experiment. However, the accuracy achieved high results with the combination of semi-realistic and not semi-realistic data, especially the Random Forest algorithm. For this reason, the F-measure values were evaluated only for the Random Forest algorithm (Fig. 7.4).

It is extremely interesting to observe how the F-measure values changed with respect to the previous experiment. Note that activities that have high F-measure

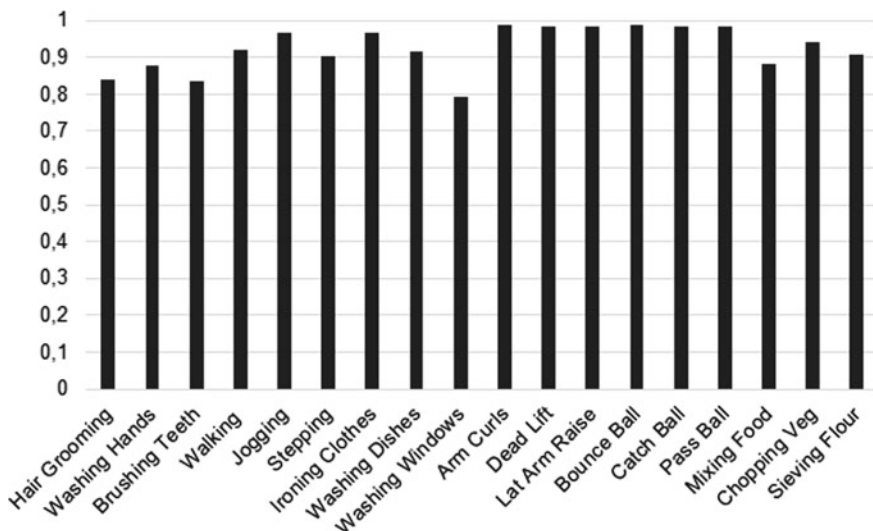


Fig. 7.4 F-measure values computed for each activity using only Random Forest classifier

values and are closest to one are increased (e.g., Sport Scenario and Exercise (Weight) Scenario).

Furthermore, this time the Bounce Ball activity did not provide a null F-measure. The combination of Dataset 1 and Dataset 2 was extremely helpful to the algorithm in both training and testing phases, simplifying the learning and recognition of such activity.

To observe more in details the recognition rate of each activity, the confusion matrix was analyzed by examining the approach with the highest accuracy that was the Random Forest one (Fig. 7.5).

The confusion matrix suggests that some specific misclassifications were more frequent than others:

- Activities of Arm Curls (j), Dead Lift (k), Lateral Arm Raise (l), Catch Ball (n) and Pass Ball (o) represent similar movements and more rhythmic repetitive movements, that are the best recognized;
- Activities containing less random movements are also not well-recognized or confused with other activities (e.g., Stepping (f) classified as Walking (d), Brushing Teeth (c) classified as Hair Grooming (a), Washing Hands (b) classified as Sieving Flour (r));
- Activities of Hair Grooming (a), Washing Hands (b), Washing Dishes (h), Washing Windows (i) and Mixing Food (p) were considered as the most challenging activities due to the natural randomness in performing these activities and for this reason they are not very well-recognized.

Both the second and third experiments were influenced by the initial choice of the eighteen activities to perform. The selected daily life activities always involved the movement of the dominant arm, but also several other parts of the body: total body activities (e.g., Walking and Catch Ball), predominantly arm-based activities (e.g., Brushing Teeth and Hair Grooming), and predominantly leg-based activities (e.g., Jogging and Stepping). Also, a different range of intensity was included: moderate

| a   | b   | c   | d   | e   | f   | g   | h   | i   | j   | k   | l   | m   | n   | o   | p   | q | r   | <-- classified as |                         |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|-----|-------------------|-------------------------|
| 135 | 1   | 9   | 0   | 0   | 0   | 0   | 0   | 2   | 15  | 0   | 1   | 0   | 0   | 0   | 0   | 1 | 2   | 0                 | a = Hair Grooming       |
| 1   | 148 | 2   | 0   | 0   | 0   | 0   | 0   | 0   | 4   | 0   | 0   | 0   | 0   | 0   | 0   | 0 | 1   | 8                 | b = Washing Hands       |
| 7   | 5   | 111 | 0   | 1   | 0   | 1   | 5   | 0   | 0   | 0   | 0   | 0   | 1   | 1   | 0   | 3 | 4   | 0                 | c = Brushing Teeth      |
| 0   | 0   | 1   | 123 | 0   | 5   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0 | 1   | 0                 | d = Walking             |
| 0   | 0   | 0   | 0   | 142 | 2   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0 | 0   | 0                 | e = Jogging             |
| 0   | 0   | 0   | 10  | 4   | 106 | 0   | 0   | 0   | 0   | 0   | 2   | 0   | 0   | 0   | 0   | 0 | 0   | 0                 | f = Stepping            |
| 0   | 0   | 0   | 0   | 0   | 0   | 147 | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1 | 0   | 0                 | g = Ironing Clothes     |
| 4   | 2   | 2   | 0   | 0   | 0   | 0   | 140 | 1   | 0   | 0   | 0   | 0   | 0   | 2   | 0   | 1 | 0   | 0                 | h = Washing Dishes      |
| 7   | 12  | 1   | 0   | 1   | 0   | 1   | 4   | 116 | 1   | 0   | 0   | 1   | 0   | 2   | 3   | 0 | 0   | 0                 | i = Washing Windows     |
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 134 | 0   | 0   | 0   | 0   | 1   | 0   | 0 | 0   | 0                 | j = Arm Curls           |
| 0   | 0   | 0   | 3   | 0   | 0   | 0   | 0   | 0   | 0   | 125 | 0   | 0   | 0   | 0   | 0   | 0 | 0   | 0                 | k = Dead Lift           |
| 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 122 | 0   | 1   | 0   | 0   | 0 | 0   | 0                 | l = Lat Arm Raise       |
| 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 154 | 0   | 0   | 0   | 0 | 0   | 0                 | m = Bounce Ball         |
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 154 | 0   | 0   | 0 | 0   | 0                 | n = Catch Ball          |
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 142 | 0   | 0 | 1   | 0                 | o = Pass Ball           |
| 0   | 1   | 0   | 0   | 0   | 0   | 7   | 2   | 7   | 0   | 0   | 0   | 0   | 0   | 0   | 108 | 1 | 1   | 0                 | p = Mixing Food         |
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0 | 112 | 1                 | q = Chopping Vegetables |
| 0   | 5   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 1 | 3   | 105               | r = Sieving Flour       |

Fig. 7.5 Confusion matrix computed for each activity using only Random Forest classifier

intensity activities (e.g., Ironing Clothes and Sieving Flour) and vigorous activities (e.g., Mixing Food and Washing Hands). As a result, some activities were confused because they were harder to distinguish and recognize using a single device placed on the wrist. For example, using a single accelerometer worn on the wrist, the orientation of the device may be similar in some activities (e.g., Walking and Stepping). Nevertheless, we can affirm that the algorithm worked very well, recognizing most of the activities.

## 7.6 Conclusions

The overall aim of this work was to undertake a study relating to data quality and its implications in the design of HAR systems. Two datasets were considered. The Dataset 1 was collected in a previous work [9] by students in an unsupervised laboratory setting. The Dataset 2 was collected by three volunteers in a semi-realistic manner getting out of the lab. However, both used a single wrist-worn accelerometer for collecting data, selecting eighteen activities to include a wide range of common daily activities. The study investigated three different sets of features (i.e., Set 1: features selected from the previous work [9]; Set 2: combination of Set 1 and additional six commonly used features; Set 3: top selected features of Set 2 using the Relief-F algorithm). The extracted features were used as inputs for the classifiers and the different classification approaches were compared in terms of accuracy. The highest accuracy for both the datasets was achieved by using features from Set 1, which gave the best result in recognition rate especially with Random Forest approach (i.e., Dataset 1: 93.63% and Dataset 2: 93.82%). Contrary to what often may be thought, this study demonstrates that a higher number of extracted features do not necessarily cause a better recognition performance for the machine learning approaches, especially if irrelevant and redundant features are added.

Furthermore, there is no doubt that there are many differences between daily life activities and activities performed in laboratory setting. We exploited the small Dataset 2 to illustrate the challenges for the machine learning approaches in the real-world applications. One approach may be effective in a laboratory setting but may have many difficulties in real-world application scenarios. The problems become evident when the accuracy of the algorithms trained and tested on the same dataset is very high and may be not completely truthful (e.g., in this specific case, the accuracy dropped by 10% up to 40% by training on pretended activities and then testing the algorithm in a realistic dataset). As a result, this study demonstrates that such algorithms should deal with activities of different nature and complexity, tackling the challenges in the real-world conditions. This means that the use of pretended activities in HAR studies should be avoided whenever possible or limited only to activities that can be reliably reproduced, as if performed out of laboratory settings. The last ones are typically those activities that do not require the manipulation of tools or objects. The combination of both two datasets was undertaken to a similar analysis. In particular, Dataset 3 included data representative of eighteen activities abounding in

inter-subject variability about how the activities were performed. Generally speaking, different activities have a different impact, unique for everyone. This is evident in a large-scale dataset collected by many individuals that produce high diversity and generalization within the dataset. This great variability allowed to learn more details about activities. Hence, in the validation phase, the algorithm had more capability to distinguish and recognize several activities.

For the third time, Random Forest achieved the highest performance. As a consequence, we can conclude that Random Forest is a very handy algorithm that easily produces a good prediction of results in a short time development. Additionally, we can say that the combination of two completely different datasets may be a good idea to build a versatile model that well-generalizes, being effective across a range of users, inputs and applications.

## 7.7 Future Works

The results of this work show the potential implications that data quality may have on HAR performance. This study would deal with the consequences of making decisions in HAR, testing some possible solutions to improve the results.

In future works, some developments and improvements may be related to different aspects, like including more volunteers performing activities, in order to enlarge the Dataset 2. In order to increase the individual's performance, the acquisition time might be shorter than 2 min for performing and recording each single activity. After repeating the same movements for 1 min with the same arm, individuals feel tired and annoyed. This might reinforce our preliminary results.

A further acquisition system might acquire a broader range of activities, with the incorporation of activities performed in real-life setting. This choice might allow the analysis of a realistic daily monitoring. A larger dataset might also involve a second person, who could help the volunteer pressing the start-stop streaming button and checking the manner to perform activity. This collaboration could reduce the delays or missing data, especially in short duration activities. Additionally, the investigation of activities might involve a higher number of accelerometers in different positions to distinguish that movements that might be similar. The placement of sensor is important in the success of algorithms in human activity recognition involving several body parts. Although one single accelerometer has proven to achieve good results, other sensing devices (e.g., gyroscopes) may be combined with accelerometers for improving the recognition performance.

## References

1. Anguita D, Ghio A, Oneto L et al (2013) A public domain dataset for human activity recognition using smartphones, pp 24–26
2. Ariani A, Koesoema AP, Soegijoko S (2017) Innovative healthcare systems for the 21st century
3. Banos O, Galvez JM, Damas M et al (2014) Window size impact in human activity recognition. *Sensors (Switzerland)* 14:6474–6499. <https://doi.org/10.3390/s140406474>
4. Bao L, Intille SS (2004) Activity recognition from user-annotated acceleration data, pp 1–17. [https://doi.org/10.1007/978-3-540-24646-6\\_1](https://doi.org/10.1007/978-3-540-24646-6_1)
5. Brox Ó (2013) The value of health care information exchange and interoperability. *Atalante* 74–83. <https://doi.org/10.1377/hlthaff.w5.10>
6. Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput Surv* 1:1–33. <https://doi.org/10.1145/2499621>
7. Chen L, Khalil I (2010) Activity recognition: approaches, practices and trends (Chap. 3)
8. Cleland I, Kikhia B, Nugent C, Boytsov A, Josef H, Synnes K et al (2013) Optimal placement of accelerometers for the detection of everyday activities, pp 9183–9200
9. Cleland I, Donnelly MP, Nugent CD et al (2018) Collection of a diverse, naturalistic and annotated dataset for wearable activity recognition, pp 674–679. <https://doi.org/10.1109/percomw.2018.8480322>
10. Cook DJ, Krishnan NC (2015) Activity learning: discovering, recognizing, and predicting human behavior from sensor data
11. Guiry JJ, P Van De Ven, Nelson J (2014) Multi-sensor fusion for enhanced contextual awareness of everyday activities with ubiquitous devices. 5687–5701. <https://doi.org/10.3390/s140305687>
12. Gupta P, Dallas T (2014) Feature selection and activity recognition system using a single triaxial accelerometer. *IEEE Trans Biomed Eng* 61:1780–1786. <https://doi.org/10.1109/TBME.2014.2307069>
13. Hall AM (1999) Correlation-based feature selection for machine learning. University of Waikato
14. John G, Kohavi R, Pflieger K (1994) Irrelevant features and the subset selection problem. *Icml* 121–129. <https://doi.org/10.1126/science.333.6044.823-c>
15. Khusainov R, Azzi D, Achumba IE, Bersch SD (2013) Real-time human ambulation, activity, and physiological monitoring: taxonomy of issues, techniques, applications, challenges and limitations. *Sensors (Switzerland)* 13:12852–12902. <https://doi.org/10.3390/s131012852>
16. Kim E, Helal S, Cook D (2010) Human activity recognition and pattern discovery. *IEEE Pervasive Comput* 9
17. Kong Y, Fu Y (2018) Action recognition and human interaction. *Hum Act Recognit Predict A Surv* 13. [https://doi.org/10.1007/978-3-319-27004-3\\_2](https://doi.org/10.1007/978-3-319-27004-3_2)
18. Kononenko I (1994) Estimating attributes: analysis and extensions of RELIEF. *Mach Learn ECML-94* 784:171–182. <https://doi.org/10.1007/3-540-57868-4>
19. Kou Z, Wu C (2018) Smartphone based operating behaviour modelling of agricultural machinery. *IFAC-PapersOnLine* 51:521–525. <https://doi.org/10.1016/j.ifacol.2018.08.156>
20. Lara OD, Labrador MA (2013) A survey on human activity recognition using wearable sensors. *IEEE Commun Surv Tutor* 15:1192–1209. <https://doi.org/10.1109/SURV.2012.110112.00192>
21. Lester J, Choudhury T, Borriello G (2006) A practical approach to recognizing physical activities, pp 1–16
22. Mannini A, Intille SS, Rosenberger M et al (2013) Activity recognition using a single accelerometer placed at the wrist or ankle. 2193–2203. <https://doi.org/10.1249/mss.0b013e31829736d6>
23. Morales J, Akopian D (2017) Physical activity recognition by smartphones, a survey. *Biocybern Biomed Eng* 37:388–400. <https://doi.org/10.1016/j.bbe.2017.04.004>
24. Morales J, Akopian D, Aгаian S (2014) Human activity recognition by smartphones regardless of device orientation human activity recognition by smartphones regardless of device orientation. <https://doi.org/10.1117/12.2043180>



25. Nakamoto K, Konishi Y, Kondo K, Ishigaki H (1999) Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *IDC 1999—1999 Information, Decis Control Data Inf Fusion Symp Signal Process Commun Symp Decis Control Symp Proc* 15:283–288. <https://doi.org/10.1109/idc.1999.754171>
26. Nugent C et al (2016) Improving the quality of user generated data sets for activity recognition. In: García C, Caballero-Gil P, Burmester M, Quesada-Arencibia A (eds) *Ubiquitous computing and ambient intelligence. IWAAL 2016, AmIHEALTH 2016, UCAMI 2016. Lecture Notes in Computer Science, vol 10070. Springer, Cham*
27. Preece SJ, Goulermas JY, Kenney LPJ et al (2009) Activity identification using body-mounted sensors—a review of classification techniques. *Physiol Meas* 30. <https://doi.org/10.1088/0967-3334/30/4/r01>
28. Randell C, Muller H (2000) Context awareness by analysing accelerometer data, pp 175–176
29. Twomey N, Diethel T, Fafoutis X et al (2018) A comprehensive study of activity recognition using accelerometers. *Informatics* 5:27. <https://doi.org/10.3390/informatics5020027>
30. Vaizman Y, Weibel NG, Lanckriet N (2017) Context recognition in-the-wild: unified model for multi-modal sensors and multi-label classification. *PACM Interact Mob Wearable Ubiquitous Technol* 1:1–22. <https://doi.org/10.1145/3161192>
31. Vavilis S, Petković M, Zannone N (2012) Impact of ICT on home healthcare. *IFIP Adv Inf Commun Technol* 386 AICT:111–122. [https://doi.org/10.1007/978-3-642-33332-3\\_11](https://doi.org/10.1007/978-3-642-33332-3_11)
32. Viet VQ, Thang HM, Choi D (2012) Balancing precision and battery drain in activity recognition on mobile phone. 1–2. <https://doi.org/10.1109/icpads.2012.108>
33. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. *Front Robot AI* 2. <https://doi.org/10.3389/frobt.2015.00028>
34. Weber DM, Kauffman RJ (2011) What drives global ICT adoption? Analysis and research directions. *Electron Commer Res Appl* 10:683–701. <https://doi.org/10.1016/j.elerap.2011.01.001>
35. Yang JY, Wang JS, Chen YP (2008) Using acceleration measurements for activity recognition: an effective learning algorithm for constructing neural classifiers. *Pattern Recognit Lett* 29(16):2213–2220. <http://linkinghub.elsevier.com/retrieve/pii/S0167865508002560>
36. Yiyan L, Fang Z, Wenhua S (2016) An hidden markov model based complex walking pattern recognition algorithm. In: 2016 fourth international conference on ubiquitous positioning, indoor navigation and location based services (UPINLBS), pp 223–229. <https://doi.org/10.1109/upinlbs.2016.7809976>



# Chapter 8

## Automated General Movement Assessment for Perinatal Stroke Screening in Infants



**Yan Gao, Yang Long, Yu Guan, Anna Basu, Jessica Baggaley and Thomas Plötz**

**Abstract** Perinatal stroke (PS) is a serious condition that often leads to life-long disability, in particular cerebral palsy (CP). Early detection and early intervention could improve motor outcome. In clinical settings, Prechtl's general movement assessment (GMA) can be used to classify infant movements using a Gestalt approach, identifying infants at high risk of abnormal motor development. Training and maintenance of assessment skills are essential and expensive for the correct use of GMA, yet many practitioners lack these skills, preventing larger-scale screening and leading to significant risks of missing affected infants. We present an automated approach to GMA, based on body-worn accelerometers and a novel sensor data analysis method—discriminative pattern discovery (DPD)—that is designed to cope with scenarios where only coarse annotations of data are available for model training. We demonstrate the effectiveness of our approach in a study with 34 newborns (21 typically

---

An extended version of this chapter was published in Proceedings of the ACM on Interactive, Mobile, wearable and Ubiquitous Technology (IMWUT), vol. 3, issue 1, article 12, March 2019 (Doi: <https://dl.acm.org/citation.cfm?doid=3323054.3314399>).

Y. Gao and Y. Long contributed equally with joint first authorship.

---

Y. Gao · Y. Long · Y. Guan

Open Lab, School of Computing, Newcastle University, Newcastle upon Tyne, UK

e-mail: [y.gao47@newcastle.ac.uk](mailto:y.gao47@newcastle.ac.uk)

Y. Long

e-mail: [yang.long@ieee.org](mailto:yang.long@ieee.org)

Y. Guan

e-mail: [yu.guan@newcastle.ac.uk](mailto:yu.guan@newcastle.ac.uk)

A. Basu

Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

e-mail: [anna.basu@newcastle.ac.uk](mailto:anna.basu@newcastle.ac.uk)

J. Baggaley

Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK

e-mail: [jess.baggaley2@newcastle.ac.uk](mailto:jess.baggaley2@newcastle.ac.uk)

T. Plötz (✉)

School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

e-mail: [thomas.ploetz@gatech.edu](mailto:thomas.ploetz@gatech.edu)

© Springer Nature Switzerland AG 2020

F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications and Networks, [https://doi.org/10.1007/978-3-030-25590-9\\_8](https://doi.org/10.1007/978-3-030-25590-9_8)

developing infants and 13 PS infants with abnormal movements). Our method is able to correctly recognise the trials with abnormal movements with at least the accuracy that is required by newly trained human annotators (75%), which is encouraging towards our ultimate goal of an automated screening system that can be used population-wide.

**Keywords** Human activity recognition · Health · Wearables · Machine learning · Prechtl's general movements assessment · Perinatal stroke

## 8.1 Introduction

Perinatal stroke (PS), i.e. a stroke occurring before or around the time of birth of an infant, is a serious concern because it can lead to conditions such as cerebral palsy (CP) with negative impact on quality of life and independence. CP is a term used to describe a group of disorders with life-long adverse effects on movement and posture due to damage incurred to the developing brain. It is the most common motor disorder in childhood, affecting around 2 in 1,000 children [1]. A common form is unilateral CP, with weakness and stiffness affecting one side of the body. Unilateral CP is often caused by PS. Around 1 in 3,500 infants sustain PS, and between 10 and 50% of infants with PS develop CP [2].

Early therapy intervention in unilateral PS is under investigation [3, 4] and may improve motor outcomes [5]. However, early intervention requires early detection of affected infants. This is challenging, as PS does not present with immediate signs of unilateral weakness, in contrast to stroke in adults; clinical features at the time of stroke are less specific and in some cases not detectable, with later emergence of motor problems over several months. Furthermore, there is no routine screening program for PS. Cranial magnetic resonance imaging (MRI) could detect PS but is costly (e.g. £1,000 per scan) [6]. It is performed if indicated based on the clinical condition, i.e. if there is a high level of suspicion of neurological abnormality. If the infant is extremely preterm, cranial ultrasound screening is usually undertaken but may miss cases of PS [7].

Another option for screening and early detection of infants at risk of motor disorders such as CP is a standardised, manual visual observation procedure according to Prechtl's general movements assessment (GMA) [8]. It is used both clinically and in research to monitor infant movements in the first several months of life with high predictive validity [9]. The GMA relies on the recognition and distinction of specific aspects of the infants' spontaneous movement repertoire during the first months of life. However, it is not a diagnostic test—it highlights infants in need of further investigation and likely intervention. The GMA requires the infant to be placed in the supine position for around 5–10 min in a quiet, alert state while being video-recorded. The video is later scored by a trained professional (see Sect. 2.1 for more details). Reliable use of the GMA requires attendance on a specific training course, followed by extensive practice to obtain and maintain the skills required for Gestalt detection

of abnormal infant movements. Many practitioners do not currently have these skills. Furthermore, the assessment is inherently subjective, and observer performance may be affected by fatigue, inter-subject variability, etc.

With the proliferation of wearable and pervasive sensing, combined with breakthroughs in machine-learning-based sensor data analysis, we hypothesise that automated GMA-based screening of infants can be realised. Our goal is to develop methods that enable population-wide, minimal effort yet accurate and objective assessments of every newborn that will lead to early detection of potential motor abnormalities. Such an automated screening procedure will not lead to fewer cases of PS but to earlier detection. The earlier PS can be detected, the earlier dedicated treatments can be administered, aiming to improve outcome. The focus of our research agenda is to develop simple, high-fidelity assessments, which are the basis for accurate and objective diagnosis.

In this chapter, we lay the foundation for the aforementioned agenda. Triaxial accelerometers are attached to each limb of typically developing infants and those with PS to record their movements. Through a machine-learning-based analysis pipeline, these sensor data are then automatically analysed with the objective of identifying infants who show signs of abnormal movements (AM) that may be indicative for PS. Sensor data are of high temporal resolution, which is of importance for the detection of even subtle indicators of PS (according to GMA). We have collected a dataset from 34 children (21 typically developing, and 13 with PS) who were observed as part of a research study. Data were collected in multiple trials each, at monthly intervals and each infant provided at least one trial between the age of 1 and 6 months. Video recordings of the trials were reviewed separately by a trained professional blinded to the clinical condition and classified (according to GMA criteria or, for age 6 months which is beyond the GMA window, by expert clinical observation) as normal or abnormal.

Strictly speaking the automated sensor data analysis resembles a human activity recognition (HAR) problem, which has been widely studied in the ubiquitous computing community including many applications in health and well-being (e.g. [10–14]). A specific challenge that is linked to GMA-related movement assessment in infants (but not uncommon in the wider health-related HAR domain, e.g. [11]) lies in the sparse annotation of the sensor data that can be used for modelling, which prevents the use of standard activity recognition pipelines such as [15], or even more contemporary deep-learning-based methods [16–20] that are notorious for relying on large amounts of annotated sample data. Only one label is available per trial (assessment session) that in itself typically has a duration of several minutes: the infant shows abnormal movements or not. In real-world clinical scenarios, it is simply not feasible to provide sample-precise annotation that would allow to model dedicated events that are linked to abnormal movements. To address this sample/annotation imbalance problem, we have developed a novel analysis framework that explicitly focuses on sparse labelling. We present a discriminative pattern discovery (DPD) framework that can suppress less informative segments for detecting behaviour patterns that corresponding to abnormal movements, which enables robust modelling that leads to correct and reliable automated assessments. Based on an experimental evaluation

using the recorded dataset, we demonstrate that these automated assessments are of comparable accuracy to those provided by human experts, which is encouraging for our overall agenda of developing screening methods that can be used at population scale. Given that wearable accelerometers are now very inexpensive and can thus be considered commodity devices, and the minimal effort needed for data recording, the proposed solution can be considered the first milestone towards our goal of generalised screening. The contributions of this chapter can be summarised as follows:

**Approach:** We developed a method for automated detection of abnormal movements (AM) in infants. Through employing inexpensive, wearable accelerometers, the assessment method is straightforward to integrate into everyday practice with minimal effort, which has the potential to significantly alleviate the cost of medical-expert-based GMA. By employing machine-learning-based sensor data analysis methods, confounding factors of manual GMA assessments can be overcome, including skill and experience (or lack thereof) of a human assessor, fatigue, inter-subject variability, subjectivity, etc.

**Dataset:** We have collected a considerable dataset of 34 infants (21 typically developing, 13 with perinatal stroke) wearing the lightweight accelerometers on their limbs during multiple assessment sessions (trials). This dataset includes realistic trial-wise expert annotations for model training and evaluation.

**Analysis Method:** In response to the challenges of the data recording and especially annotation procedure, we have developed a DPD framework for modelling that effectively tackles the trial-wise classification tasks as a weak labelling problem. We employ a novel kernel-based algorithm to improve model generalisation.

**Model Performance:** Based on the recorded dataset, we demonstrate that our automated analysis method can surpass the pass mark (75% in accuracy) for the GMA examination required for a human to practise this method.

## 8.2 Background

### 8.2.1 *Clinical Routine for Diagnosing Perinatal Stroke*

Around 10–50% of infants who suffer from PS develop CP [2], due to damage to those parts of the brain that control movement, balance and posture. Through observation of spontaneous movements of an infant, and through neurological examinations supported by cranial imaging, the clinical diagnosis of CP can be made. Although cranial imaging was used to provide a definitive gold standard by which to classify the infants as having had PS or not, definitive imaging with MRI is costly (around £1,000 per scan for infants in the UK), and requires sedation or general anaesthetic in some cases, which has an associated risk [21]. Therefore, MRI-based brain imaging of infants is not a routine screening procedure, even for preterm infants, though it is used as diagnostic evaluation of symptomatic infants.

On the other hand, movement monitoring for infants is widely applied in predicting impairments in neuro-motor development, and various qualitative methods have been proposed to assess the quality of motion patterns for infants [8]. A more often used approach for identifying infants at high risk is the use of Prechtl's general movements assessment (GMA) [8]. Specifically, the term "general movements" describes infant movements in the quiet alert state based on a Gestalt perception of normal versus abnormal movements that naturally encompasses factors such as the complexity, fluency and variation of the movements, and from around 3 months, the emergence of "fidgety" movements in the typically developing infant. GMA can be done within several minutes, and it is purely observational. However, it cannot classify infants as having PS or not: it can merely describe infant movements as being abnormal or normal. GMA provides prognostic information if used correctly and thus infants identified on a clinical basis as having abnormal GMA would be followed up and investigated (including brain imaging as appropriate). The main difficulty with this assessment is that it requires extensive observer training and practice to retain the skills required for Gestalt detection of abnormal movements—many practitioners do not have these skills.

The clinical classification process according to GMA is age-dependent, because the characteristics of spontaneous infant movements evolve with age. Furthermore, the clinical classification of general movements is based on the Gestalt impression of the quality of movement over the whole segment of the video recording viewed, of a supine infant in the quiet, alert state, paying attention to complexity, fluency and variation. GMs are scored as normal or abnormal based on this impression. GMs in the first two months of life (corrected for prematurity, i.e. from the expected date of delivery onwards) are predominantly "writhing" in nature, with relatively slow, fluent, elegant, complex, variable rotational movements. The three subclassifications of types of abnormal GMs in this age group are "poor repertoire" GMs (lack of complexity and variety in the movement repertoire), "cramped-synchronised" GMs (with periods of synchronous whole body muscle contraction and relaxation) and "chaotic GMs" (extreme lack of movement fluency as well as large amplitude of movements). From age 6 weeks onwards, "fidgety" movements may begin to emerge; they peak in their frequency of occurrence at around 12–16 weeks and are typically observed until around 20 weeks though may persist a few weeks longer. Fidgety movements are smaller in amplitude, and higher in speed, than writhing movements. It is abnormal for fidgety movements to be absent, or to be present but very large amplitude and excessively fast and jerky. Thus, for infants aged 3–5 months (corrected for prematurity), the two subclassifications of abnormal types of GMs are "absent fidgety movements" and "abnormal fidgety movements". For infants aged 6 months, voluntary movements are normally predominant; movements for infants aged 6 months were therefore classified as normal or abnormal by an experienced observer based on whether appropriate voluntary movements were observed.

## 8.2.2 *Movement Assessment with Wearables*

Recently, wearable technologies have been used for automated capturing and subsequent analysis of spontaneous movements of infants with a view to reproducing the GM classification without the need for a clinical observer [22]. In [23], a clinical tool based on accelerometers was used to assess motion patterns for preterm infants, and it was demonstrated that using accelerometry data was a reliable way to evaluate the characteristics of movement disorders [24]. Singh et al. developed a system that leverages accelerometers on detecting features related to CP through analysing abnormal movements in premature infants [25]. Fan et al. developed a Markov-model-based technique that can recognise gestures from accelerometers. They demonstrated that by treating instantaneous machine learning classification values as observations and explicitly modelling duration, the recognition rate of abnormal movements can be improved [26].

From the perspective of ubiquitous computing, the aforementioned techniques fall in the field of human activity recognition (HAR), which plays a major role in computational behaviour analysis, with applications in, for example, daily life monitoring [27], medical diagnosis [27, 28], sports tracking and coaching [13], skill assessment [29, 30], health and well-being assessment [11], to name but a few. In general, the goal of HAR is to automatically recognise what a person is doing, and when. Over the years, a multitude of methods have been developed, which is not surprising given that activity recognition constitutes one of the main pillars of ubiquitous computing contributing to automated context inference [31]. It is beyond the scope of this chapter to give a detailed overview of activity recognition techniques, much of which can now even be considered somewhat common sense in the field (for excellent surveys on the field cf., e.g. [10, 15, 32, 33]).

The predominant analysis approach is based on transforming raw sensor data into a sequence of isolated analysis frames by employing a sliding-window procedure [15]. This sliding-window procedure needs to be optimised very carefully to determine appropriate length of the analysis frames and overlap between subsequent frames. Typically, these parameters are optimised heuristically and globally for all activities of interest (even though recently variations with optimised window lengths for each activity of interest have been developed [34]). Subsequently, for every analysis frame, features are extracted that are fed into classification backends that automatically decide upon the labels, that is class (activity) associations, of the portions of sensor data that are captured by the individual frames [35–37]. This standard activity recognition chain follows the supervised modelling approach, and fine-grained ground truth annotation is required for model training and optimisation: every sensor reading (sample) needs to be annotated, which is typically realised through manual observation and annotation on the recording timeline. Implicitly, this annotation is translated into a sample-wise labelling, but it requires a human observer to pay very close attention to the annotation process itself. For real-world deployment scenarios, such time-consuming annotation procedures—especially for the sake of collecting annotated sample data for method development—are often unfeasible as they

interfere with the primary task such as providing care. Instead, more coarse annotation schemes are employed such as providing assessments on a session or trial level, as pursued in our application scenario. On a technical level, this trial-wise annotation translates into a weak supervision problem with few labels only for many sensor readings, which introduces ambiguity that rules out standard supervised model training procedures. Our analysis method effectively addresses this weak supervision problem through automatically discovering the subtle discriminative patterns within the recordings of complete analysis sessions that can differentiate abnormal movements (AM) from typical developing (TD) infants.

## 8.3 Methodology

### 8.3.1 Data Acquisition

*Participants.* We recruited 21 typically developing (TD) infants and 13 infants with PS. Ethical approval was obtained prior to the study (West of Scotland Research Ethics Committee, reference number 15/WS/0129), and all necessary research governance processes were followed. Infants were excluded if they: (i) had any additional significant medical diagnoses, such as severe visual impairment, which could render outcomes uninterpretable; (ii) had radiological evidence of significant bilateral intra cerebral pathology or that only the occipital, prefrontal or temporal lobes (i.e. non-motor areas) were affected; (iii) had ongoing involvement in another study which would likely interfere with this study. Descriptive statistics of the involved participants' assessments can be found in Table 8.1.

*Data Collection.* All participants underwent a ten-minute recording while lying supine at monthly intervals of their birthday from term age, or term equivalent age for preterms, for six months. To record movement during the supine recording, infants wore ankle and wrist straps, specifically designed for use with Axivity WAX9 inertial movement unit (IMUs [3]) which include accelerometer, gyroscope and

**Table 8.1** Descriptive statistics of assessments by age, diagnosis and gestation (#: number of trials)

| Month | Typical developing |          |        | Perinatal stroke |          |        |
|-------|--------------------|----------|--------|------------------|----------|--------|
|       | #Term              | #Preterm | #Total | #Term            | #Preterm | #Total |
| 1     | 6                  | 3        | 9      | 2                | 6        | 8      |
| 2     | 11                 | 4        | 15     | 5                | 6        | 11     |
| 3     | 12                 | 4        | 16     | 5                | 7        | 12     |
| 4     | 14                 | 4        | 18     | 5                | 5        | 10     |
| 5     | 14                 | 3        | 17     | 5                | 4        | 9      |
| 6     | 16                 | 4        | 20     | 5                | 5        | 10     |
| Total | 73                 | 22       | 95     | 27               | 33       | 60     |

magnetometer) recordings. In this work, only triaxial accelerometer data were used. The straps were made of four brightly coloured cottons (yellow, blue, green and red) with a sealable IMU pouch and an individual piece of Velcro hooks sewn onto one end of the strap. The inside of the strap was made of pile felt which sticks to the Velcro hooks therefore creating a secure, adjustable comfortable fit for each infant.

The IMUs are lightweight (approx. 10 g) with recorded acceleration at 100 Hz in three axes ( $x$ ,  $y$  and  $z$ ). The WAX9 devices were switched on and sealed into waterproof bags prior to being inserted into the pocket of the straps, with the Axivity logo arrow face up and pointing into the strap's pouch. To synchronise the IMUs, short impulse forces, created by shaking all the IMUs at once, were used to create data markers [38, 39]. At the end of the recordings, the IMUs were subjected to a second set of short impulse forces at the same time in front of the video camera to provide a second set of data markers. The whole process was also recorded by a camera for annotation purposes. By following this protocol, we inserted markers into both the sensor recordings and the video footage, which enabled time alignments between the data streams. The data recording procedure is straightforward because it does not interfere with normal infant handling routine. The sensors can easily be slipped into the pouches. The tight fit and the clear indications of orientation alignment greatly reduce the risk of misalignment or inserting the sensors with the wrong orientation. The pouches are integrated into colour-coded straps that easily fit the limbs of newborns. With their fully adjustable fit and range of sizes, they are easy to put on. It takes only a few seconds to attach the sensor to a limb, a routine that can easily be integrated into general handling of the infant, e.g. during play.

We also recorded some meta information for this dataset. For example, 'head position' was recorded based on the position of the infant's head as either 'left', 'right' or 'midline'. 'Left' and 'Right' were determined as the infant having turning their head over  $30^\circ$  away from the midline in either direction. In order to constrain the requirements of our approach to what is feasible in real-life application scenarios, we did not use the head position information, even though it may provide important context. The other meta information we recorded is the quality of the trials. Unqualified trials were excluded for this study, such as when the infant had rolled on to their front, or was crying/sneezing/sick/coughing, etc. As a result, we obtained a number of continuous trials each 4–10 min in length. A professional trained in GMA observed each video and classified infant movements as normal (TD) or abnormal (AM).

### 8.3.2 Data Pre-processing

After removing some unqualified/corrupted trials, we had 161 validated continuous trials (some longer trials were broken into shorter ones). Each trial contained four sets of synchronised sequences that corresponded to the four accelerometers fixed on the infant's four limbs. Each accelerometer dataset (collected from a limb) consisted of three axis data ( $x$ ,  $y$  and  $z$ ) and their signal vector magnitude as a complementary



dimension, i.e. four channels (per limb). The trial length was roughly 4–10 min, i.e. 24 k–60 k of samples. Each trial was labelled indicating whether the data came from a typical developing (TD) infant or an infant with abnormal movements (AM). In total, there were 64 positive trials (AM) and 97 negative trials (TD).

For data processing, we employed the sliding-window approach without any window/frame overlapping. We further concatenated each 1s window (i.e. 100 samples) of 16 channels raw data (i.e. from four limbs) into a  $16 \times 100$ -dimensional vector. As a result, each trial was segmented into a number of windows, while each window corresponded to 1s movement records. We aimed to identify windows of abnormal movements (wAM), and most importantly also to make an overall decision (i.e. AM/TD) for the whole trial, which corresponded to the Gestalt nature of GMA. To further improve the efficiency and remove redundancy, we conducted PCA dimension reduction to project each window of raw signals into a low-dimensional feature space for further processing.

### 8.3.3 Challenges

*Weak Labels.* GMA aims at classifying normal/abnormal movements for complete trials/sessions, and it normally relies on intensive observations of rigid and characterised body motions to match a pre-defined score sheet. Existing automated approaches [26] rely on expensive expert annotations, such as the cramped-synchronised general movements (CSGM) [40]. Such a paradigm requires sample-wise annotations and is therefore less feasible for the purpose of building up a large-scale training set. In fact, the number of qualified annotators is also very limited. One must undergo special training before being able to provide reliable annotation. The process is time-consuming, and the annotator needs to mark the start and end time on the sequences when CSGM occurs, which often suffers from severe inter-subject variability. In our case, we only have data with realistic, hence sparse trial-wise annotation (TD vs. AM). We need to train a system that not only can classify an unseen trial as either TD or AM, but also be able to indicate the moments when patterns of abnormality occurred (i.e. to identify wAM).

*Generalisation.* Even though we underwent an extensive and time-consuming data collection process (including participant recruitment, logistics for the session, actual data recording and cleaning, and data annotation), the resulting dataset is, strictly speaking in machine learning terms, still small-scale, which poses substantial pressure and constraints on the analysis methods. Therefore, the second challenge is to prevent the model from overfitting by improving its generalisation capabilities for small-scale training datasets.

### 8.3.4 Automated GMA Through Analysing Accelerometer Data

*Problem Formalisation.* Although we have trial-wise annotation, we are facing a weak label problem in modelling the system. That is, there is no label information for each window/frame within the trials, which may serve as important clues for final classification decision. In this work, we define (unseen) window labels as:

1. wAM: window of abnormal movement, which are unlikely performed by TD infants.
2. wTD: window of movements of a typical developing infant that cannot easily be performed by AM infants.
3. wNM: window of neutral movement, which are the common movements that can be easily performed by both AM/TD infants.

Based on this, within any query trial, we can assess each window independently and aggregate all corresponding scores for the final trial-level decision. Figure 8.1 illustrates the general framework of our system, visualising the two-stage assessment process, window-wise classification (wAM/wTD/wNM), followed by overall trial-wise classification (AM/TD). In what follows we will provide the details of the developed analysis approach.

Raw signals of each window can be represented using low-dimensional PCA features  $\mathbf{x} \in \mathbb{R}^d$  ( $d = 100$  in this work), and in this case the trial  $X_i$  can be expressed as a collection of independent windows/instances such that  $X_i = \{\mathbf{x}_1, \dots, \mathbf{x}_{n_i}\}$  with a trial-wise label  $y_i \in \{0, 1\}$  indicating AM (1) or TD (0). In the inference stage, for query trial  $\hat{X}$ , we aim to predict its trial-wise label  $\hat{y} \in \{1, 0\}$  by accumulating all the window-wise predictions:

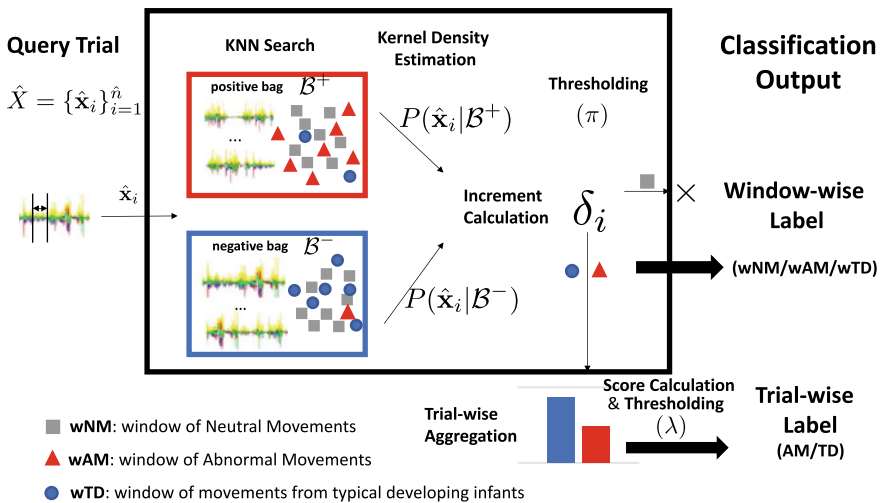


Fig. 8.1 Overview of the proposed system for automated general movement assessment

$$\hat{y} = \arg \max_{y \in \{0,1\}} \prod_{i=1}^{\hat{n}} P(y|\hat{\mathbf{x}}_i). \quad (1)$$

However, owing to the lack of window-wise annotation, it is not appropriate to apply existing machine learning algorithms directly to this task. To address this issue, in this work we define three intermediate patterns (wAM, wTD and wNM), which indicate discriminative information (wAM, wTD) to be aggregated, or redundant information (wNM) to be filtered out. Since the discriminative movement patterns may only occur sporadically for the whole trial, it is essential to reduce the redundant neutral movement information. To achieve this, we propose a framework named DPD, which can automatically classify each window/instance within the trial, yielding improved trial-wise classification.

*Discriminative Pattern Discovery.* In the field of machine learning, there are several ways of dealing with the aforementioned weak labelling problem. One of the most popular ways is to treat it as a Multiple Instance (MI) learning problem [41]. In MI, instead of receiving a label for each data instance, labels are given through bags, i.e. trials in our case. Specifically, MI assumes a bag is positive if it contains at least one positive instance, and it is negative otherwise. Such an assumption is similar to our problem since trials of AM infants can be modelled as a positive bag including at least one positive instance (e.g. wAM) and many negative instances. However, such presence-based MI approaches cannot be applied directly to our problem due to some key violations to its assumption. For example, MI assumes that a negative bag cannot have positive instances, whereas, in our case, even TD infants (negative trials) may show occasional ‘abnormal’ behaviour at times. As an alternative, the count-based Generalised Multiple Instance (GMI) learning was previously proposed [42]. It shows a certain level of tolerance to the presence of positive (negative) instances in negative (positive) bags. It requires a maximum as well as a minimum number of instances of a certain concept in a bag. For example, a positive bag may include at least a pre-defined number of positive instances, while at most a pre-defined number of negative instances [42]. We adopt and extend the idea of GMI.

The key idea of our DPD procedure is to compute the *increment* of each instance and classify it into three classes rather than two. Specifically, for each instance instead of classifying it as positive or negative before counting the frequencies for a bag-level decision, we identify its movement patterns into three pre-defined classes (or three distributions)—wAM (or  $\mathcal{A}$ ), wTD (or  $\mathcal{R}$ ), wNM (or  $\mathcal{O}$ )—followed by an aggregation on the corresponding increments/scores. Our assumption is that all the instances can be drawn from these three distributions  $\mathcal{A}$ ,  $\mathcal{R}$  and  $\mathcal{O}$ . The classification based on DPD depends on the “soft” score proportion rather than a “hard” presence/absence criteria in conventional MI and GMI approaches. For example, TD infants can have brief periods of apparently abnormal movement. Only if the frequency and occurrence (in terms of increment/score) of these observations exceed the thresholds, the trial is then considered as AM. Accordingly, we design two DPD rules:

1. Rule  $\pi$ : Detecting instances drawn from the distribution of  $\mathcal{O}$  (i.e. wNM) and suppress their consequential weights.
2. Rule  $\lambda$ : Learn a proportion of instances drawn from  $\mathcal{A}$  (i.e. wAM),  $\mathcal{R}$  (i.e. wTD) to differentiate AM/TD trials.

Formally, we put all instances of positive (AM) and negative (TD) trials into two bags  $\{\mathcal{B}^+, \mathcal{B}^-\}$ . For each instance in a query trial  $\hat{\mathbf{x}}_i \in \hat{X}$ , we are interested classifying it as wAM (i.e. from  $\mathcal{A}$ ), wTD (i.e. from  $\mathcal{R}$ ), or wNM (i.e. from  $\mathcal{O}$ ). This is measured by the increment function  $\delta_i$  defined using log-odds:

$$\delta_i = \log \frac{P(\mathcal{B}^+ | \hat{\mathbf{x}}_i)}{P(\mathcal{B}^- | \hat{\mathbf{x}}_i)}, \quad (2)$$

which can measure to what extent an instance can differentiate the AM/TD trials. Given  $\delta_i$ , we can perform instance-level classification via rule  $\pi$ :

$$\hat{\mathbf{x}}_i \in \begin{cases} \mathcal{A}, & \text{if } \delta_i > \pi \\ \mathcal{R}, & \text{if } \delta_i < -\pi \\ \mathcal{O}, & \text{otherwise.} \end{cases} \quad (3)$$

Note that instances drawn from  $\mathcal{A}$  or  $\mathcal{R}$  provide discriminative information, while instances drawn from  $\mathcal{O}$  reflect neutral movements which are redundant. Intuitively, if discriminative instances are outnumbered by neutral ones, that may lead to unreliable predictions. By applying rule  $\pi$ , we can classify  $\hat{\mathbf{x}}_i$  into the aforementioned three classes and apply a simple masking operation such that  $\hat{h}_i = 0$  if  $\hat{\mathbf{x}}_i \in \mathcal{O}$ , and  $\hat{h}_i = 1$  otherwise. Based on this masking operation, we can easily filter out the nondiscriminative movement patterns (wNM) for the query trial, leaving discriminative ones for the trial-wise assessment. That is, the overall trial  $\hat{X}$  can be assessed by aggregating all discriminative patterns, and rule  $\lambda$  can be applied for the final trial-wise classification (TD/AM).

$$\hat{y} = \begin{cases} 1, & \text{if } \frac{1}{\sum_{i=1}^{\hat{n}} \hat{h}_i} \sum_{i=1}^{\hat{n}} \hat{h}_i \delta_i > \lambda \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

*Kernel Generalisation.* For machine-learning-based systems, generalisation capability to unseen environments is of great importance. In our case, considering the fragility of the population and the constraints imposed on assessment procedures on infants, collecting large amounts of training data is expensive. So far, our dataset contains 161 validated trials that vary with regard to age, gender, terms, etc. These factors result in large variability in movement patterns and thus may degrade the model performance for unseen environments. In this case, sophisticated algorithms may suffer from overfitting problems with poor generalisation. Our solution for this problem follows the GMI paradigm [41, 42] which assumes that each instance is drawn from smoothed distributions by learning effective kernel functions.

A typical way is to compress the training set into a very small codebook by unsupervised clustering. Each cluster can be viewed as a representative pattern that summarises surrounding samples into a compact kernel space so that the large variability can be reduced. However, such approaches often suffer from the *quantisation problem*, i.e. the inherent inaccuracy of representation when mapping the original data to a few representatives. For example, we have roughly 70,000 instances which shall be clustered into, e.g. 200–1,000 representative patterns, which inevitably leads to loss in precision of the representation. As will be shown in the experiments, although quantifying specific instances to clusters may mitigate variability, we may also sacrifice discriminative information substantially. In addition, existing clustering algorithms (e.g. K-means) often suffer from randomness and low efficiency. These problems motivate us to design a better generalisation so as to circumvent the quantisation issue. Now we aim to model the increment  $\delta_i$  for query instances, and via Bayesian theorem, Eq. (2) can be reformulated as:

$$\delta_i = \log \frac{P(\hat{\mathbf{x}}_i|\mathcal{B}^+)P(\mathcal{B}^+)}{P(\hat{\mathbf{x}}_i|\mathcal{B}^-)P(\mathcal{B}^-)} = \log P(\hat{\mathbf{x}}_i|\mathcal{B}^+) - \log P(\hat{\mathbf{x}}_i|\mathcal{B}^-) + c, \quad (5)$$

where  $c = \log \frac{P(\mathcal{B}^+)}{P(\mathcal{B}^-)}$  is a prior constant that can be estimated by the proportion between numbers of AM and TD trials.

With this we now aim to estimate the density  $P(\hat{\mathbf{x}}_i|\mathcal{B})$  for both bags. We employ an online algorithm based on nonparametric probability density estimation [43]. The primary idea is to only consider task-relevant instances at the test time instead of blind quantisation at the training stage. Given a query instance, the algorithm first searches for a number of similar instances (e.g. via k-nearest neighbour method) in a certain training bag, which can be used to smooth the noise and variability accordingly. This process is supported by the theory of long-tail characteristic of high-dimensional features, i.e. instances that are far away from the query data in the feature space often make less contribution to the density estimation. Specifically, we estimate the kernel density of the query data by taking the integral of these surrounding relevant instances, such that:

$$P(\hat{\mathbf{x}}_i|\mathcal{B}) = \frac{1}{k} \sum_{j=1}^k \mathcal{K}(\hat{\mathbf{x}}_i - NN_j) \quad (6)$$

where  $K : \chi \times \chi \rightarrow \mathbb{R}$  is a Parzen kernel function that is non-negative and integrates to 1. Without loss of generality, we use a Gaussian kernel:  $\mathcal{K}(\hat{\mathbf{x}}_i - NN_j) = \exp(-\hat{\mathbf{x}}_i - NN_j^2)$ . Note that  $NN_j$  is the  $j$ th of top  $k$ -nearest neighbour search results of  $\mathbf{x}_i$  for bag  $\mathcal{B}$ . With Eqs. (5) and (6), the increment function  $\delta_i$  is:

$$\delta_i = \log \frac{1}{k} \sum_{j=1}^k \exp(-\hat{\mathbf{x}}_i - NN_j^{+2}) - \log \frac{1}{k} \sum_{j=1}^k \exp(-\hat{\mathbf{x}}_i - NN_j^{-2}) + c. \quad (7)$$

As mentioned before,  $c$  is a constant and can be merged into the  $\lambda$  rule (for the trial-wise thresholding) and thus does not need explicit computation. The overall algorithm is summarised in Algorithm 1.

**Algorithm 1: Discriminative Pattern Discovery with Kernel Generalisation**

**Input:** Positive and Negative Bags  $\mathcal{B}^+$ ,  $\mathcal{B}^-$ ; Query trial  $\hat{X} = \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n\}$ ; Empirical Hyper-parameters  $k$ ,  $\pi$ , and  $\lambda$ ;

**Output:** Prediction  $\hat{y}$ .

$h = 0$ ;

**for** all query instances  $\hat{\mathbf{x}}_i \in \hat{X}$  **do**

    Computer  $k$ -Nearest Neighbours of  $\hat{\mathbf{x}}_i$  in  $\mathcal{B}^+$ ;

    Computer  $k$ -Nearest Neighbours of  $\hat{\mathbf{x}}_i$  in  $\mathcal{B}^-$ ;

    Computer the increment  $\delta_i$  using Eq. (1);

**if**  $|\delta_i| < \pi$  **then**

$\hat{h}_i = 0$ ;

**else**

$\hat{h}_i = 1$ ;

$h \leftarrow h + 1$ ;

**end**

$\Delta \leftarrow \Delta + \hat{h}_i \delta_i$ ;

**end**

**if**  $\frac{1}{h} \Delta > \lambda$  **then**

$\hat{y} = 1$ ;

**else**

$\hat{y} = 0$ ;

**end**

## 8.4 Experimental Evaluation

Our evaluation design is inspired by the GM Trust Course on the Pechtl assessment of general movements. We compare our approach to a list of baseline methods as the ablation experiment. The evaluation is based on our collected dataset as introduced in Sect. 3.1.

### 8.4.1 Settings

Our experiments are based on tenfold random split (149/12) cross-validation. That is, every fold contains 12 trials as test (see below). Over the remaining training trials, we further use tenfold (137/12) cross-validation to find the optimal hyper-parameters. This setting focuses on theoretical model comparison and thus considers the participant as one type of covariate factor together with age, head position, etc.

### 8.4.2 Baselines

We compare the proposed DPD method to a range of alternative, conventional machine learning methods. Despite popularity, existing deep learning techniques (e.g. [16–20]) are not applicable to this problem because these methods need training data with sample/window-wise annotations, which are not available in our case. The baselines are as follows:

**KNN:** Since our kernel embedding is based on  $k$ -nearest neighbour (KNN) search, an intuitive baseline is KNN itself. Specifically, we first search  $k$ -nearest neighbours of every query instance  $\hat{\mathbf{x}}_i$  in the training set (i.e. AM/TD trials). The instance is classified as AM or TD depending on the  $k$ -nearest instances’ trial-wise labels. Trial-wise classification is performed via majority voting on classified instances.

**SVM:** Here all the instances in the AM trial are presumed to share the same label—AM—despite a fairly large portion of them may not present significant abnormal movements. Considering SVM is based on learning the support vectors between two distinctive distributions, we concern whether such a mechanism can be readily applied. We implement the SVM with a Gaussian kernel.

**GMI-GEN:** One of our key arguments for *Kernel Generalisation* is that clustering-based approaches may degrade the performance due to quantisation and loss of original discriminative information. In order to validate this hypothesis, we run Generalised Multiple Instance-Generative (GMI-GEN) baseline experiments. The GMI-GEN model is implemented by initially quantising the training set into a compact set of representative clusters using K-means, while we keep other details exactly the same as our proposed method.

**Ours (no-DPD):** The key difference of our method to conventional MI approaches is that we classify instances into three classes and suppress the redundant wNM. Such a procedure refers to the hyper-parameter  $\pi$  in Algorithm 1. In this baseline, we take all of the instances (also including the redundant wNM) into account without applying the  $\pi$  rule. All other parameters are kept exactly the same as for our proposed method.

**Prechtl’s Standard:** To achieve a basic certificate, a human observer needs to correctly classify at least 75% (in our case 9/12) of randomly selected AM/TD trials. Our evaluation simulates such a “random-12” setup as it is used for clinician training and assessment, and thus, the results are directly interpretable for practical scenarios.

### 8.4.3 Evaluation Metrics

We use accuracy as the main metric to compare different baselines to our approaches. Due to the (slightly) imbalanced distribution of AM and TD trials (see Table 8.1), we adopt metrics based on true positive (TP), false positive (FP), true negative (TN) and false negative (FN) classifications as follows:

**Accuracy:** Number of correctly classified trials, normalised over all trials.

**Sensitivity (Recall):**  $TP/(TP + FN)$  measures the proportion of positives that are correctly identified as AM.

**False-positive rate:**  $FP/(FP + TN)$  shows to what extent TD are confused as AM.

**Specificity:**  $TN/(FP + TN)$ , also known as true-negative rate, measures the proportion of negatives that are correctly identified as TD.

**Precision:**  $TP/(TP + FP)$ , fraction of AM trials among overall positive predictions.

We also provide confusion matrices and receiver operating characteristic (ROC) curves for further discussion.

### 8.4.4 Results

Table 8.2 shows the results for different baselines and our method. It can be seen that our method consistently outperforms other methods with the highest overall accuracy (80%). It is worth noting that the proposed automated assessment method is able to correctly classify AM with at least the accuracy that is required by trained human annotators (75%) to pass the GM examinations. Our model can therefore provide an objective, automatically generated reference for clinical purposes. Next, we analyse the contributions of our approach through comparing it to baselines.

The first two baselines have lowest recognition performance. KNN predicts all of the query trials as AM, whereas SVM goes to the other extreme with all TD predictions. We ascribe the failure of KNN to poor generalisation because test instances of new trials may be very dissimilar to the training trials. Similarly, the performance of the SVMs shows its low generalisation—despite the fact that we employ a Gaussian

**Table 8.2** Classification accuracies for tenfold cross-validation [means; std. ( $\pm$ )]

| Method        | Accuracy          | Sensitivity (recall) | False-positive rate | Specificity       | Precision         |
|---------------|-------------------|----------------------|---------------------|-------------------|-------------------|
| KNN           | 0.22(0.12)        | 1.00(0.00)           | 1.00(0.00)          | 0.00(0.00)        | 0.22(0.12)        |
| SVM           | 0.79(0.11)        | 0.00(0.00)           | 0.00(0.00)          | 1.00(0.00)        | –                 |
| GMI-GEN       | 0.32(0.17)        | 0.73(0.33)           | 0.80(0.23)          | 0.20(0.23)        | 0.20(0.12)        |
| Ours (no-DPD) | 0.70(0.10)        | <b>0.88(0.16)</b>    | 0.32(0.13)          | 0.68(0.13)        | 0.43(0.20)        |
| Ours          | <b>0.80(0.13)</b> | 0.70(0.35)           | <b>0.13(0.11)</b>   | <b>0.87(0.11)</b> | <b>0.57(0.27)</b> |

Bold represents standard deviation



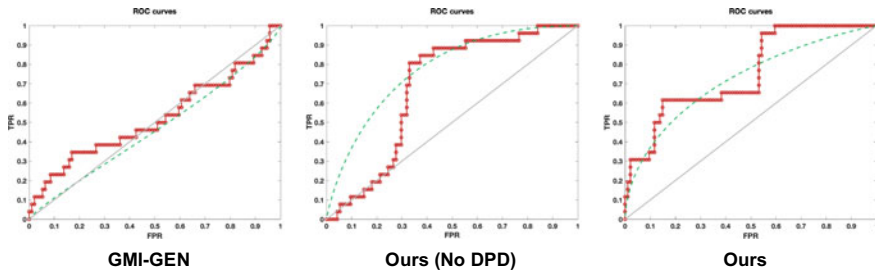


Fig. 8.2 ROC curves of key baselines

kernel that explicitly aims to capture local distributions for better generalisation. Another reason for the SVM’s failure may be the weak labelling problem, and in the experiments, we presume all the instance labels share the same one with the trial. The experimental results suggest that traditional machine learning methods cannot be used directly in our scenario, since they cannot cope well with the weak labelling problem.

GMI-GEN is a variant of our method, and it initially performs clustering for compact representation. We notice that its recall is reasonable, yet the false-positive rate is very high (0.8). Generally, GMI-GEN is significantly worse than our approach, since the discriminative information may be lost during the clustering process. The second-best approach (i.e. no-DPD) is the variant of our algorithm that does not eliminate contribution from the instances drawn from  $\mathcal{O}$  (i.e. wNM). The accuracy is only 10% lower than our approach, and the recall is even higher. We observe that its main weakness is the specificity and false-positive rate (19% lower and 19% higher, respectively). These results suggest the importance of suppressing redundant wNM via DPD’s  $\pi$  rule.

Figure 8.2 shows the ROC curves of the best three approaches (dashed curves denote smoothed trend lines). GMI-GEN performs closely to random guess but gets better when the recall is low (higher TPR/FPR). The overall area under curve of no-DPD is slightly higher than that of DPD. However, the DPD approach converges to perfect TPR earlier ( $\sim 0.6$  compared to 0.85 of No-DPD), which means that this approach is more stable and can provide more reliable predictions with lower costs. Confusion matrices in Fig. 8.3 for GMI-GEN, no-DPD and DPD provide class-based insights that allow to assess the trade-off between AM and TD predictions.

## 8.5 Summary and Conclusion

Stroke is a serious event that can lead to substantially adverse outcomes. While it is relatively straightforward to diagnose a stroke in adults, a stroke before or right after birth is much harder to detect. Such Perinatal Strokes (PS) are not uncommon but, unfortunately, many cases remain undiagnosed for too long with the loss of the

| Predicted<br>Ground Truth | Predicted   |             |
|---------------------------|-------------|-------------|
|                           | AM          | TD          |
| AM                        | 0.73 ± 0.33 | 0.27 ± 0.33 |
| TD                        | 0.80 ± 0.23 | 0.20 ± 0.23 |

GMI-GEN

| Predicted<br>Ground Truth | Predicted   |             |
|---------------------------|-------------|-------------|
|                           | AM          | TD          |
| AM                        | 0.88 ± 0.16 | 0.12 ± 0.16 |
| TD                        | 0.32 ± 0.13 | 0.68 ± 0.13 |

Ours (No DPD)

| Predicted<br>Ground Truth | Predicted   |           |
|---------------------------|-------------|-----------|
|                           | AM          | TD        |
| AM                        | 0.70 ± 0.35 | 0.30 ± 0. |
| TD                        | 0.13 ± 0.11 | 0.87 ± 0. |

Ours

**Fig. 8.3** Confusion matrices of key baselines [means; standard deviations ( $\pm$ )]

potential for early intervention aiming to improve outcomes.. A large percentage of infants affected with PS will develop conditions such as Cerebral Palsy that have life-long impact. In this chapter we presented an automated assessment system, which detects abnormal movements (AM) of infants with Perinatal Stroke. Movement data are captured through miniaturised, inexpensive inertial measurement units and analysed through a novel, machine-learning-based sensor data analysis method. This data analysis approach effectively tackles a weak label problem in which only very few and rather coarse ground truth annotations are provided for model training. In practice, clinicians classify trials of, for example, five minutes duration with regards to whether they consider the observed infant as typically developing or showing abnormal movements. The developed Discriminative Pattern Discovery (DPD) method automatically detects relevant patterns and bootstraps effective classification models based on these.

We evaluated the effectiveness of the developed framework in a deployment study. Data were recorded and analysed for a total of 34 infants (21 typically developing, 13 with clinically diagnosed PS—each confirmed through MRI), each at various steps in their development after birth. For this dataset our automated assessment system was able to correctly discriminate between TD and AM, with a higher accuracy than a GMA-trained clinician would need to achieve in order to be able to pass their course examination. These are very encouraging results because they demonstrate that automated PS screening is possible. Due to the low costs of the sensing hardware and the minimal effort to use these, larger-scale uptake is realistic.

Perinatal Strokes are often difficult to detect in the first months of life, yet can have very serious, adverse consequences. Early detection of infants with Perinatal Stroke is essential to provide timely support and medical input to mitigate against adverse outcomes. In this chapter we have demonstrated that it is possible to use body-worn inertial measurement units and novel machine learning methods for sensor data analysis to automatically distinguish between the abnormal movements of a group of infants with Perinatal Stroke, and the normal movements of control infants. Through a rigorous evaluation of our method in a cohort of infants, who either had been diagnosed with Perinatal Stroke or were typically developing, we have laid the foundation for a screening tool that can potentially be used at population scale with

minimal effort to enable early stage recognition of abnormal movements in young infants. Our method is straightforward to apply, inexpensive, and reliable with regards to the accuracy of analysis results.

**Acknowledgments** This work was supported jointly by Medical Research Council (MRC, UK) Innovation Fellowship (MR/S003916/1), Engineering and Physical Sciences Research Council (EPSRC, UK) Project DERC: Digital Economy Research Centre (EP/M023001/1) and National Institute of Health Research (NIHR, UK) Career Development Fellowship (CDF-2013-06-001)(AB). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care (DHSC, UK).

## References

1. Johnson A (2002) Prevalence and characteristics of children with cerebral palsy in Europe. *Dev Med Child Neurol* 44(9):633–640
2. Nelson KB (2007) Perinatal ischemic stroke. *Stroke* 38(2 Suppl):742–745
3. Basu AP et al (2018) Feasibility trial of an early therapy in perinatal stroke (eTIPS). *BMC Neurol* 18(1):102
4. Basu AP et al (2017) Participatory design in the development of an early therapy intervention for perinatal stroke. *BMC Pediatr* 17(1):33
5. Basu AP (2014) Early intervention after perinatal stroke: opportunities and challenges. *Dev Med Child Neurol* 56(6):516–521
6. Edwards AD et al (2018) Effect of MRI on preterm infants and their families: a randomised trial with nested diagnostic and economic evaluation. *Arch Dis Child Fetal Neonatal Ed* 103(1):F15–F21
7. Cowan F et al (2005) Does cranial ultrasound imaging identify arterial cerebral infarction in term neonates? *Arch Dis Child Fetal Neonatal Ed* 90(3):F252–F256
8. Einspieler C, Prechtl HF (2005) Prechtl’s assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. *Ment Retard Dev Disabil Res Rev* 11(1):61–67
9. Kwong AKL et al (2018) Predictive validity of spontaneous early infant movement for later cerebral palsy: a systematic review. *Dev Med Child Neurol* 60(5):480–489
10. Avci A et al (2010) Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey. In: 23th international conference on architecture of computing systems, 2010
11. Hammerla NY et al (2015) PD disease state assessment in naturalistic environments using deep learning. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, 2015. AAAI Press, Austin, Texas, pp 1742–1748
12. Hoey J et al (2011) Rapid specification and automated generation of prompting systems to assist people with dementia. *J Pervasive Mob Comput* 7(3):299–318
13. Kranz M et al (2013) The mobile fitness coach: towards individualized skill assessment using personalized mobile devices. *J Pervasive Mob Comput* 9(2):203–215
14. Plötz T, Moynihan P, Pham C, Olivier P (2011) Activity recognition and healthier food preparation. In: Chen NCL, Biswas J, Hoey J (eds) Activity recognition in pervasive intelligent environments. Atlantis Press, Atlantis Ambient and Pervasive Intelligence
15. Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. *J ACM Comput Surv* 46(3):1–33
16. Guan Y, Ploetz T (2017) Ensembles of deep LSTM learners for activity recognition using wearables. *J Proc ACM Interact Mob Wearable Ubiquitous Technol* 1(2):1–28

17. Hammerla NY et al (2016) Deep, convolutional, and recurrent models for human activity recognition using wearables. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, 2016. AAAI Press, New York, New York, USA, pp 1533–1540
18. Ordóñez FJ, Roggen D (2016) Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16(1):115
19. Yang JB et al (2015) Deep convolutional neural networks on multichannel time series for human activity recognition. In: Proceedings of the 24th international conference on artificial intelligence, 2015. AAAI Press, Buenos Aires, Argentina, pp 3995–4001
20. Zeng M et al (2014) Convolutional neural networks for human activity recognition using mobile sensors. In: 6th International conference on mobile computing, applications and services
21. Salerno S et al (2018) Is MRI imaging in pediatric age totally safe? A critical appraisal. *Radiol Med* 123(9):695–702
22. Marcroft C et al (2014) Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. *Front Neurol* 5:284
23. Gravem D et al (2012) Assessment of infant movement with a compact wireless accelerometer system. *J Med Dev* 6:2
24. Heinze F et al (2010) Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy. *Med Biol Eng Comput* 48(8):765–772
25. Singh M, Patterson DJ (2010) Involuntary gesture recognition for predicting cerebral palsy in high-risk infants. In: International symposium on wearable computers (ISWC), 2010
26. Fan M et al (2012) Augmenting gesture recognition with erlang-cox models to identify neurological disorders in premature babies. In: Proceedings of the 2012 ACM conference on ubiquitous computing (UbiComp)
27. Plöetz T, Hammerla NY, Olivier P (2011) Feature learning for activity recognition in ubiquitous computing. In: Proceedings of the twenty-second international joint conference on artificial intelligence, vol 2. AAAI Press, Barcelona, Catalonia, Spain, pp 1729–1734
28. Bachlin M et al (2010) Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Trans Inf Technol Biomed* 14(2):436–446
29. Khan A et al (2015) Beyond activity recognition: skill assessment from accelerometer data. In: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, 2015. ACM, Osaka, Japan, pp 1155–1166
30. Ladha C et al (2013) ClimbAX: skill assessment for climbing enthusiasts. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing, 2013. ACM, Zurich, Switzerland, pp 235–244
31. Abowd GD (2012) What next, ubicomp? celebrating an intellectual disappearing act. In: Proceedings of the 2012 ACM conference on ubiquitous computing, 2012. ACM, Pittsburgh, Pennsylvania, pp 31–40
32. Chen L et al (2012) Sensor-based activity recognition. *IEEE Trans Syst Man Cybern Part C (Appl Rev)* 42(6):790–808
33. Plötz T, Guan Y (2018) Deep learning for human activity recognition in mobile computing. *Computer* 51(5):50–59
34. Li H et al (2018) On specialized window lengths and detector based human activity recognition. In: Proceedings of the 2018 ACM international symposium on wearable computers, 2018. ACM, Singapore, Singapore, pp 68–71
35. Figo D et al (2010) Preprocessing techniques for context recognition from accelerometer data. *Pers Ubiquit Comput* 14(7):645–662
36. Hammerla NY et al (2013) On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In: Proceedings of the 2013 international symposium on wearable computers, 2013. ACM, Zurich, Switzerland, pp 65–68
37. Kwon H et al (2018) Adding structural characteristics to distribution-based accelerometer representations for activity recognition using wearables. In: Proceedings of the 2018 ACM international symposium on wearable computers, 2018. ACM, Singapore, Singapore, pp 72–75

38. Ploetz T et al (2012) Automatic synchronization of wearable sensors and video-cameras for ground truth annotation —a practical approach. In: 2012 16th international symposium on wearable computers
39. Hsu WY et al (2012) Effects of repetitive transcranial magnetic stimulation on motor functions in patients with stroke: a meta-analysis. *Stroke* 43(7):1849–1857
40. Ferrari F et al (2002) Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy. *Arch Pediatr Adolesc Med* 156(5):460–467
41. Dietterich TG et al (1997) Solving the multiple instance problem with axis-parallel rectangles. *J Artif Intell* 89(1–2):31–71
42. Weidmann N, Frank E, Pfahringer B (2003) A two-level learning method for generalized multi-instance problems. In: Proceedings of the 14th European conference on machine learning, 2003. Springer, Cavtat-Dubrovnik, Croatia, pp 468–479
43. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, Hoboken

**Part III**  
**User Needs and Personalisation**

# Chapter 9

## User-Centered Design in Defining and Developing Health and Well-Being ICT Solutions



Nikolaos Liappas, José G. Teriús-Padrón, Rebeca I. García-Betances, María Fernanda Cabrera-Umpiérrez and María Teresa Arredondo

**Abstract** Implementing emerging technologies is a complex task which requires time, precision, and organization. The definition of a logical structure for classifying and organizing complex information during the design process of a technological solution provides flexibility while the process becomes more prescriptive. One of the most used methods in the ICT field is the user-centered design methodology (UCD). UCD methodology situates the final user as the cornerstone of the research and development process since the success or failure of a technological solution will depend on users' acceptance. The following chapter presents experiences, best practices and lessons learned applying UCD methodology in different European projects from several years of work conducted at LifeSTech group from UPM, in areas such as: chronic diseases management, accessibility, and cognitive rehabilitation. Specifically, the chapter explains how the UCD methodology was applied during the different stages of the design and development process for different domains and use cases.

**Keywords** User-centered design · Active assisted living · ICT · Well-being

### 9.1 Introduction

The development of ICT applications for the healthcare domain involves and requires the cooperation and interaction of multidisciplinary stakeholders (e.g., health professionals, end users, engineers, etc.) in order to guarantee the successful and acceptability of the developed applications. Because of the complexity of this interaction, different methodologies and frameworks are used to help ICT teams to include these actors during all the phases of the design and development process.

---

N. Liappas · J. G. Teriús-Padrón · R. I. García-Betances (✉) · M. F. Cabrera-Umpiérrez · M. T. Arredondo  
Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain  
e-mail: [rgarcia@lst.tfo.upm.es](mailto:rgarcia@lst.tfo.upm.es)

N. Liappas  
e-mail: [nliappas@lst.tfo.upm.es](mailto:nliappas@lst.tfo.upm.es)

User-centered design is a design philosophy and process that considers the users and their requirements as the most important factor to improve the quality of designs. This process relies hardly on user's continuous feedback during all phases of the design and development process of the solution [1]. The term 'User-Centered Design' was presented by Donald Norman in the publication of a co-authored book entitled: *User-Centered System Design: New Perspectives on Human-Computer Interaction* [2]. In this book, Norman et al. adapted and re-defined the participatory design concept into the user-centered design philosophy. Later, according to Abras et al. [3] Norman built further on the UCD and recognizes the needs of the user and the usability of the design, presenting four basic suggestions on how a design should be. These conditions are: (1) make it easy to determine what actions are possible at any moment; (2) make things visible, including the conceptual model of the system, the alternative actions, and the results of actions; (3) make it easy to evaluate the current state of the system; (4) follow natural mappings between intentions and the required actions, between actions and the resulting effect, and between the information that is visible and the interpretation of the system state.

Initially, UCD included several areas of basic and applied research such as: cognitive and social psychology, linguistics, mathematics, computer science, engineering, human factors and ergonomics, socio-technical systems design, scientific management, industrial and occupational psychology, human relations and organizational behavior [4]. Since 1992 UCD has been used also in the healthcare domain, addressing different focus areas such as elderly care, cognitive care, chronic diseases care, etc., [5]. It has been used as a method for the design and development of healthcare delivery systems such as mobile and web care devices and applications, emergency systems, self-management and decision support systems.

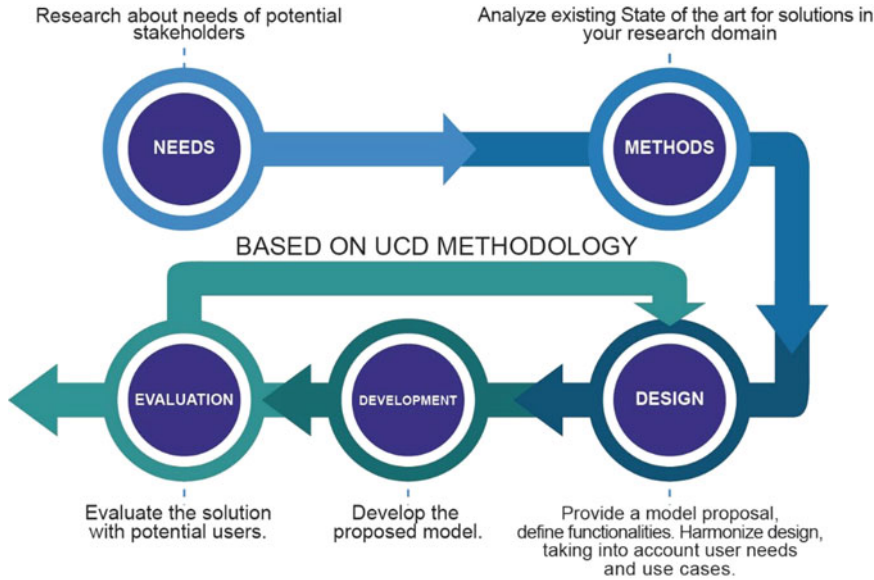
The most important phases of UCD when applied in relevant healthcare applications in domains like chronic diseases management [8] and cognitive rehabilitation [6, 7] are:

- **Needs and technical requirements identification:** include the collection of users' needs and their transformation into technological requirements. This could be done using different methods such as: interviews, observations, questionnaires, workshops, and focus groups.
- **Definition of use cases:** refers to the description, during the design process, of the user actions and all possible interactions with the system.
- **Validations:** is the phase when users test the proposed solution and give their feedback. This process could involve different assessment methods such as: heuristic evaluation, usability tests, end user evaluation, pre- and post-survey, among others.

UCD phases can be adapted to the design and the develop process of an ICT solution as shown in Fig. 9.1, where five stages of the development are described as follow:

- **Needs:** Research about the needs of the potential users of the solution.
- **Methods:** Analyze the existing technologies in the domain of your solution.
- **Design:** Provide a proposal of the solutions, define functionalities based on the user needs.





**Fig. 9.1** An example of the user-centered design methodology in health care

- Development: The proposed solution.
- Evaluation: Perform with potential users and redefine if needed.

In this chapter, we present the UCD methodology applied in different ICT healthcare-related and research-based projects hosted by our research group. Because of the diversity of these projects, we present them divided based on their focus domain. Specifically, we present some important aspects and ways of applying UCD for different use cases, adaptations that should be taken into consideration when applying UCD in certain scenarios, and the outcomes, most important insights, lessons learned, and recommendations gathered from these experiences that could help researchers and developers to deliver better solutions.

## 9.2 Applying UCD in Different Domains

In this section, we present our experiences, best practices, and recommendations applying UCD in high impact ICT European projects, from the Life Supporting Technologies group, in areas such as: chronic diseases management, accessibility and cognitive rehabilitation, and training.

### **9.2.1 Chronic Diseases Self-management**

According to the World Health Organization, chronic diseases are responsible for almost 70% of all deaths worldwide. In addition, 82% of the 16 million people who died before reaching age 70 due to chronic diseases take place in low- and middle-income countries. These types of diseases have shown to produce devastating health consequences for individuals, families, and communities [9].

The self-management of chronic diseases is one of the most used techniques nowadays because it gives the patient the tools to manage and improve their condition. The management skills provided by a self-management program include problem solving, decision making, resource utilization, establishing a patient–caregiver partnership, and undertake actions [6].

At the Life Supporting Technologies group, we have applied the UCD methodology specifically in technological projects for patients with chronic diseases such as: cardiovascular diseases, diabetes and Parkinson. All these projects aimed to design and develop an ICT solution that, not only could help patients to manage their disease in the best possible way, allowing them to improve their quality of life, but also provide useful information to healthcare professionals to manage patients' health conditions.

#### **9.2.1.1 Experiences Applying UCD in ICT Solutions for Chronic Diseases Management**

In this section, we present the experiences and lessons learned during the phases of technical requirements' definitions, use cases, and validation, in chronic disease management projects for patients with diseases such as diabetes, cardiovascular, and Parkinson.

Technical requirements' definitions:

The major goal of this phase is to identify the relevant set of medical and technical requirements that will be the starting point for developing the specifications of the overall architecture. Conducting research, indagating and agreeing on requirements during the design phase, can help during the development of the project facilitating the project planning.

To determine the technical requirements of an ICT solution, we first must consider that not only is it necessary to do a theoretical investigation of the specific kind of chronic disease and its treatment, but also to receive useful information from potential stakeholders (e.g., physicians, patients, caregivers, nurses, users, etc.). For example, in a study conducted for the HeartCycle project, in which we developed a personal health system for cardiovascular diseases management, a total of 26 patients were interviewed at the beginning of the design phase. The interviews followed a protocol

that included a presentation of the project concept, a conversation on a storyboard that showed some of the potential functionalities, and a final questionnaire with a scoring sheet. The results were analyzed in order to identify patients' major problems and needs in specific workshops that involve experts and health professionals [10].

Based on our experience, we have been able to detect that, although each chronic disease is different in symptoms, the evolution and treatment to follow, the needs and goals of patients, doctors and caregivers are similar. First, patients seek out to have a tool that helps them to manage their health and at the same time provide educational content, reminders, physical activity monitoring, feeding guide, emergency interventions and allow them to have greater independence and control on the disease leading to an improvement of their well-being. Second, doctors and caregivers seek for a tool in which they can evaluate the evolution of their patients, their treatment adherence and maintain contact with them remotely as well.

Furthermore, other aspects to be considered in the later development phase are: (1) any solution targeted to be used by patients should be easy to use; and (2) there was a clear need to add patient education and motivation strategies into the solutions [10].

Use cases:

The definition of use cases for the management of patients with chronic diseases does not have much difference from what the theory suggests, which consists of the description of the system focused on low-level user actions including system response [11]. However, we can point out that for developments that seek to provide a complete solution that gives the necessary tools to both the patient and the healthcare professionals, different cases such as: intervention in case of emergency, management and intervention for taking medication, exercise management, diet guide, medical patient communication, etc., must be included to ensure better health management of the chronic disease.

We also recommend not ruling out any use case, even if it is considered obvious since we are talking about scenarios in where the health of the patients could be affected or where useful interventions could help them. Based on this assumption, in an initial phase, everything must be considered. For example, in the METABO project, which consisted in the development of a system to assist diabetic patients in controlling their metabolic disorders, a total of 33 use cases were clearly defined to guarantee the complete functionality of the solution and meet the expectations of all the potential users. These definitions include, as an example, a simple task as user login, the possibility to have an error during the login process and how to solve the issue [12]. This sometimes is considered evident for when developing ICT solution, but not consider it from the beginning could produce design and definition problems during the development and validation phases in some cases and in particular with specific type of users (e.g., elderly people).

## Validations:

The validation process is the key element in every ICT development process as the users are the ones that determine the usability and reliability of the solution and its quality and capacity to meet their needs. Throughout our experience, we have applied different validation methods that we describe below.

In the METABO project [13], we applied the accepted standards DIN EN ISO 9241 and ISO 3407 as the foundation of the validation process. According to these standards, usability can be measured as efficacy, efficiency, and satisfaction. A total of 36 Type 1 diabetes patients tested the solution. The usability problems were reported to the development teams and have been addressed by them following an iterative process of development. As result of this validation 43 usability issues were identified, managed, and solved.

Alternately, in the HeartCycle Project, we have applied questionnaires, combining semantic differentials and Likert scales to evaluate the usability and acceptability of the system. Each interview itself included three main parts: (1) a general introduction of the project and purpose of the test; (2) questions focused on general motivation factors; and (3) the interview with a short introduction and demonstration of the application. In the third part, the patients were asked to use the application performing specific tasks and fill in a questionnaire with their impressions. The outcomes of this validation led to a refinement of the global design and the implementation of an improved application. This upgrade took into consideration all the received suggestions and specifically addressed the aspects that obtained worse acceptance rates during the validations [14].

Finally, in the PD manager Project, which developed an m-health ecosystem for Parkinson's disease management by using non-intrusive mobile and wearable devices, the acceptability and utility of the PD manager system were assessed by comparing it with traditional practices of using a patient/career diary for the management of the disease. The validation phase was performed following a non-blinded parallel two-group randomized controlled pilot study for a period of two weeks [15]. A total number of 200 patients and 200 caregivers were enrolled and tested the solution. After the test period, interviews were made to collect their opinions to assess the acceptance, ease of use, usefulness of information of the PD manager system and the specific section of the solution related to "Symptoms Diary".

In the previous examples, several validation procedures were applied, all of them widely used by ICT researchers and development teams. However, it is important to point out that the key element of the validation phase in these kinds of solutions is to select the validation technique that best suits the needs of the different stakeholders involved and the goals of the specific study. Following this, we can objectively detect the weak points that will ultimately determine the success or failure of the solution.

## 9.2.2 Accessibility

Accessibility, in terms of human–computer interaction, refers to the computer accessibility and the capability of a computing system to be accessible to everyone, regardless the impairments, age, assistance needed, or disabilities people might have. In this domain, several use cases have been implemented through different projects and UCD has been used extensively. Thus, we present our findings and the insights we consider valuable when we applied UCD to projects dealing with accessibility issues.

### 9.2.2.1 Utilizing UCD with Virtual User Modeling to Create Accessible Application Interfaces, the Case Study of VERITAS

VERITAS presented and delivered an adaptation of UCD to serve a large portion of users with different needs such as hearing and/or visual disabilities, cognitive impairments, mobility issues, etc., or elderly people with problems related to their health conditions and the age. The motivation for this adaptation is that the application of UCD with the patient-centric approach [16] has not been explored thoroughly and the end products or services are not well adapted to the final users. The patient-centric approach refers to the way a healthcare system involves the experts together with patients and their families to adjust choices to patients' needs and inclinations. Furthermore, it incorporates the conveyance of explicit education and bolster patients need to settle on these choices and take an interest in their very own health care.

This intersection between the patient-centric approach and the UCD leads to accessible products that finally could satisfy the end users by meeting their special needs. It is important to address the accessibility issues to bridge the gap between software development methodologies and UCD. While this task is very challenging, VERITAS addressed it with the help of interaction paths. We described the interaction path by defining all the possible states that a user can go through in the system and within its graphical user interface (GUI). The solution implemented by VERITAS is described below. The procedure is to associate the interaction paths and tasks models in order to bring off scenario files [17, 18].

During the modeling and design phase, the goal-oriented design (G-OD) [12] was chosen as an approach to focus mainly on the goals of the users and define detailed the use cases and scenarios. The methodology follows iteration circles within these states: research, modeling, requirements, framework, refinement, and support. At first, we defined the goals in the research phase, then we modeled the goals with the “personas” concept in order to represent them effectively, and later, in the requirements phase, we designed the scenarios able to describe the complete context and the conditions of use. The next phase is the framework phase in which we start to design the GUI. To do so, first, we chose the available interaction devices, designed all the possible screens, and created the key paths that describe the most frequent tasks the user can perform in the system. The key paths are important, and they represent the best-case scenarios that can describe all actions with minimal steps, by means while the system

operates normally and with no human errors. In the next phase on refinement, we defined all the possible secondary tasks that can derive from the best-case scenarios.

In the field of medical applications, it is critical for the systems to be able to react in real time. Also, the users should be able to have the ability to choose their preferred way of presentation of the content according to their preferences and their specific technical knowledge. Although the interaction paths can describe effectively all the actions in one system, sometimes they disregard some important fundamentals: they remain static and they cannot react in real time to user inputs (e.g., answers to questionnaires). In order to bypass this, dynamic key paths enable the interaction workflows to meet the needs of the users in real time. Delivering a more personalized solution before the support phase while following the G-OD methodology is achieved by the following steps [17]:

- Step 1: In the refinement phase, the GUI is initiated and designed to meet the specific needs. In this step, we record the interaction sessions for creating scenarios to be used in the next step.
- Step 2: In this step, we connect the previous scenarios with the simulation models. These models are defined in the processing phase and they describe all the objects in the GUI. Then, the task models are generated, and they map and relate to the user events corresponding to the UI elements. Thus, for every task, different events are associated with the recorded images of the previous step 1. This process is very important as provides the users' activity on every task and deliver for each event all the interactions involved.
- Step 3: The task models created in the previous step 2 were exported in order to be reproduced in the GUI with different virtual user models. At this step and having created the virtual user models, we evaluated the GUI with specific groups of people each one represented from one or more virtual user models. In that manner, we evaluated the interactions and addressed all the accessibility issues. Finally, the development team applied the final changes after having implemented all the virtual models to represent all groups of users.

This approach found to be very informative as we discovered at early stages many accessibility problems of the GUI, before the development phase. Hence, the design of accessible and personal healthcare solutions can take advantage of all these insights, on an early phase, in order to deliver more adapted solutions to specific segment of users. The evaluation of all the tools and the processes followed by VERITAS demonstrated that there is a good correlation between the final outcomes of the virtual modeling with UCD approach and the real users. Finally, while tested with real users, more problems occurred but there was enough correspondence to the results of the virtual users so that this approach should be adopted by other developers or actors as well.

### 9.2.2.2 Applying Holistic UCD in ICT Health Desktop, Mobile and Web Applications, the Case Study of AEGIS

Aegis examined the third-generation access techniques and the possibility to exploit them in order to build mainstream and accessible applications including desktop and mobile applications and utilizing the rich Internet. Open accessibility framework (OAF) was developed and is open access to everyone in order to provide guidelines regarding the design and development of accessible ICT solutions. Through a holistic UCD adaptation, AEGIS explored user needs and interaction models for several user groups such as users with visual, hearing, and cognitive problems and provided insights for developers as well [19].

Because of the diversity of the applications developed, different use cases implemented on this project. All the use cases are designed to serve people with disabilities and the following groups of people are identified:

1. Cognitive impairment users
2. Hearing impairment users
3. Speech impairment users
4. Blind and low-vision users
5. Motor impairment users.

In those target groups, various health ICT applications have been developed with the UCD applied, on the domain of accessibility and they consist of desktop applications, web applications, and mobile applications. Hence, the results and lessons learned from this project provide valuable contribution to the reader and the scientific community.

The first steps were to initiate the UCD process and involve all the main stakeholders such as experts in the field, developers and users with disabilities at this early stage to have plenty of time for future modifications and adaptations. A consensus was defined and agreed to the following elements: user involvement at very early stages as high priority due to the diversity of the user groups; everything users can see, hear, and touch to be designed together with a multidisciplinary team.

The UCD plan followed four phases: thoroughly analysis of all the users to collect the needs of the users and other relevant insights; definition of requirements and use cases by translating the previous needs to a more meaningful representation; initiation of the development through the conceptual designs utilizing a co-design approach; test and evaluate the prototype with end users.

After the application of this procedure, we present the main outcomes and lessons learned through this feedback gained through a participatory design method.

Summarizing the experiences and lessons learned through holistic UCD methodology in ICT applications for the specific target groups mentioned before [20, 21]:

- Avoid all in one solution, as the requirements divert on the user groups. For example, blind users had the desire to avoid the screens and the hearing-impaired users could not adapt to this solution.

- Developers should develop with the help of the users and never release a version without the approval of at least one expert user.
- Each iteration cycle is done only after the previous iteration solved all the identified issues.
- Specific features are implemented only if they are asked by the users; developers should avoid thinking possible features without being experts on the field.
- Full hardware specifications must be given in the testers on each cycle to avoid causing problems with compatibilities.
- Developers should use or integrate already existing elements and software to gain time while respecting the copyright policy.
- People with disabilities and elderly usually lack the knowledge of assistive technologies and confuse the outcomes in the piloting phase. To avoid that, experts should be involved and provide their valuable feedback.

Evaluation performed through a framework implemented within the project in order to validate all three markets of the applications developed: rich Internet, mobile and desktop applications. Because of the UCD applicability in all stages as mentioned before, the validation techniques include all the related stakeholders and professionals while they take place in four different countries and six sites feeding with feedback the technical teams on each phase.

Because of the different applications developed in this project, different evaluation categories have identified: technical validation, human factor assessment, impact assessment, and socio-economic assessment of all the AEGIS applications. The testing sites performed the trials with the end users assessing the human factors while the more technical evaluations performed in the developers' sites. However, the remaining impact and socio-economic assessment was conducted only with the feedback of the previous technical and human factors' assessment.

For a better understanding the technical assessment examined the validity of the system, its performance, the quality offered and the accessibility issues including all the involved stakeholders. The human factors assessment considered to examine the acceptance of the system and how is it affecting the users and their lives. The impact assessment was conducted to examine if the users were satisfied and if they improved their quality of life. At the same time, it assessed the new possibilities to enter the market and deliver future business solutions in health care. Only experts were involved in this phase. Finally, the socio-economic assessment considered all the economical possibilities to pay the products performing and strengths and weaknesses analyses.

The pilots on each country performed the evaluation in three phases [22] in order to provide adequate feedback to the developers and deliver a satisfactory project. These three phases consist of: (a) test the initial conceptual design with simulations and mock-ups utilizing the Wizard of Oz technique; (b) creation of interfaces of early prototypes and for the initial testing with specific test cases; (c) perform full trials and tests with the application demonstrators with the final users and with experts involved.



### ***9.2.3 Cognitive Rehabilitation and Training***

People with cognitive impairments face various problems related to their cognition, memory, and learning capabilities. Cognitive rehabilitation refers to programs and applications tailored for individuals able to help them restore their normal cognitive functioning or to compensate deficits [23]. Computer-assisted types of rehabilitation and training have shown evidence that can be effective for individuals who suffered from a stroke, a brain trauma, or have a cognitive decline.

#### **9.2.3.1 Providing Detailed Use Case Scenarios Based on UCD and Literature Review, the Case Study of InLife**

The project focused on elderly with cognitive impairments. The solution adopted by InLife delivered ICT interoperable and personalized applications able to prolong the independent living, promote the socialization, and control the health of the people with cognitive impairments [24].

This project followed a slightly different approach to design the use cases using the UCD methodology [25]. Firstly, a detailed review was done in all previous funding calls and EU projects related to the same target user, seniors with cognitive impairments. Secondly, insights and the state of the art of several projects but for the same target groups were provided, in addition to a review of papers associated with relevant existing technologies. Furthermore, because of the division of the project in different pilots running in different regions, local workshops were held on each region in order to instantiate the findings of the user needs in terms of service functionality and variations in different locations. Finally, the outputs of the workshops combined with the previous literature analysis created a very detailed “personas” scenario that can describe the overall solution and service to potential users with great accuracy. Through this methodology, the main user requirements were exported and used to create the “personas” scenarios before the development.

To evaluate the results of the previous methodology, the project used mainly interviews and questionnaires because of the diversity of the stakeholders involved in the process. Also, to homogenize all the information from all the partners, an ad hoc template was developed and used to gather all the information in quantitative and qualitative categories and to perform various analyses on those data. After the data analysis, it was acknowledged that the extraction of user needs had been focusing on two kinds of primary users, people with cognitive impairments and healthy older people, while people living alone and wanting to maintain their independence was the majority.

Defining use cases found to be more time-consuming but effective and slightly different from the existing approaches due to its mixed elements. Using this approach, we were able to identify specific groups of people at a higher degree of abstraction and classify them better according to their needs and their associated environments.

To conclude, this approach of defining use cases found to be more sluggish but effective and slightly different from the existing approaches due to its mixed elements. The outcome is that this approach can identify specific groups of people at a higher degree of abstract, classify them better according to their needs and associated environments and provide detailed use cases.

### **9.2.3.2 Co-creation of a Hybrid UCD with Stakeholders, Products and IoT, the Case Study of ACTIVAGE**

It is worth mentioning the application of UCD and relevant insights in large-scale projects with pilots currently running simultaneously in different countries. For this reason, we present you in this section the project ACTIVAGE [26]. The main objective is the implementation of a reference framework for smart living for aging well solutions encompassing the use of Internet of Things.

In the context of UCD methodology, each deployment site (DS) adapted its own UCD method. Nonetheless, we can find common elements that can give insights for future integration of UCD in large pilots. All the pilots followed classical methods to gather the required information about the environment, stakeholders and the needs of users, professionals and caregivers and other involved stakeholders. These methods include documentary analysis, open questionnaires, structure and semi-structure interviews and focus groups.

The pilot of Madrid DS is focusing on the early detection and prevention of cognitive decline and the falling risk of elderly people in their living environment. The hybrid solution adopted by Madrid DS combines UCD with technological development approaches while exploiting the IoT and the Smart Cities paradigms. The initial results of this adaptation of the co-creation framework revealed us already on the first iterations' detailed needs and scenarios of the users and guided us to design improved technologically interventions incorporating the IoT paradigm.

The co-creation framework [26, 27] is being used specifically in the Madrid DS. Figure 9.2 presents the framework which is a combined UCD and IoT paradigm in order to identify the user needs, contextualize them to requirements, and deliver them as a service solution (provided by the demand side) combined with technology (provided by the supplier side) to the stakeholders.

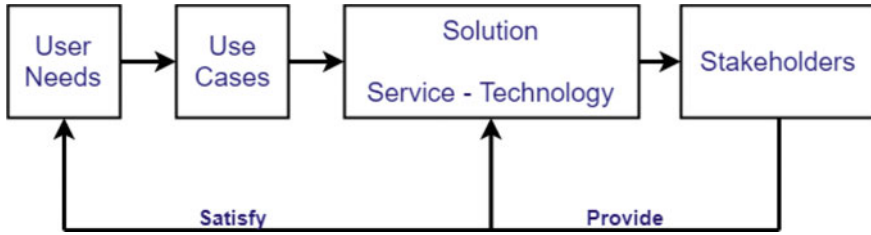


Fig. 9.2 Co-creation framework [27]

For a comprehensive understanding in the framework (Fig. 9.2): The number of the iterations is dependent on the discovery of new services and implementation, thus the desirable satisfaction of the stakeholders. In the first iteration, the previous project inputs are used to extract the user needs and couple them with the existing literature in order to design proper use case scenarios. Then, these use case scenarios are contrasted with end users and different stakeholders (e.g., elderly, caregivers). Continuing, the solution is proposed by combining the service and technology and delivered to the appropriate stakeholder for the next phase. In the second iteration, we re-pass all the elements but viewing it from a different angle. At this phase, we examine the values that we can extract based on a technological perspective and on the IoT paradigm incorporated in our proposed solution. We define and construct all the services consolidating the results on each iteration, and we utilize a waterfall generic model to assist us on the technical development with these phases: analysis, adaptation, and assessment. Finally, after each iteration, the solution is tested with the users and given the feedback we iterate until we reach a satisfactory service for the users.

This method has been useful up now to define the services on our pilot. The defined solutions have been tested in our living laboratory technically. However, because this is an on-going task, the evaluation will occur at a later phase. One of the key requirements to achieve this satisfaction and keep improving our services is the interoperability. Exploiting and expanding interoperability on all levels offers flexibility, enhances the solution, and allows reusability and scalability of the solutions.

To conclude, this methodology has been used as our core framework to design interventions on large scale projects. This enabled us possibilities to adapt it to the new emerging paradigms such as IoT and smart cities. Deriving values from these paradigms within UCD pushes the technological boundaries, optimizes the solutions and their acceptability.

**9.2.3.3 Utilizing UCD in All Phases to Implement a Satisfactory Product, the Case Study of Smart4MD**

The SMART4MD project develops and tests a health application specifically tailored to people with mild dementia. The SMART4MD application that is currently being

tested by users was developed for tablets. The trials will last two years [28]. In the Smart4MD project, the principles for Agile software development and inclusion of the user are influencing the whole design and adaptation process. A systematic review of earlier studies in the field has showed that user involvement in the design process has improved the usefulness and acceptability of the applications.

In this case, the UCD is applied before the development and the project initiation. The first step included semi-structured interviews and focus groups (first phase) with the end users, people with mild dementia, their careers and the associated professionals. The purpose of this was to get the most important insights at first, in order to understand in depth, the needs of these people and at a later stage to define properly the requirements for the solution. Based on this information, the first prototype was designed and exposed to the second stage of interviews. At that stage, the target was to explore extra undiscovered problems related to usability, user experience, and personalization to this group of people. The steps followed are typical and can be seen below:

1. Preparation to identify and address important tasks and questions, ethical issues, information sheet, informed consent, guide for the moderator, etc.
2. Focus groups/interviews/observations with patients and careers (inclusion, why, how, when, whom, how many, structure.)
3. Interviews with healthcare professionals (why, how, when, whom, how many, structure.)

Finding of both phases showed important outcomes as they revealed information about the familiarity of technology with those people, past experiences with technology, content preferable for them, potential usage/agreement of the solution's features, requirements and usage motivation and ethics related issues. UCD played an important role in revealing all this information and making the solution adapted and adopted by people with mild dementia.

More specifically, the processed followed can be summarized in the following Fig. 9.3. This UCD process can be used as a guide for future implementation when working with specific people such as people with mild cognitive impairments and mild dementia.

### 9.3 Highlights and Conclusion

Through this chapter, we have shown our experiences and lessons learned over the years dedicated to research and development of ICT solutions applied in multiple domains, using user-centered design methodology. This methodology puts the end user in the center of attention since the success of innovation depends on them. So, considering their experience, needs and expectations are fundamental during the whole process.

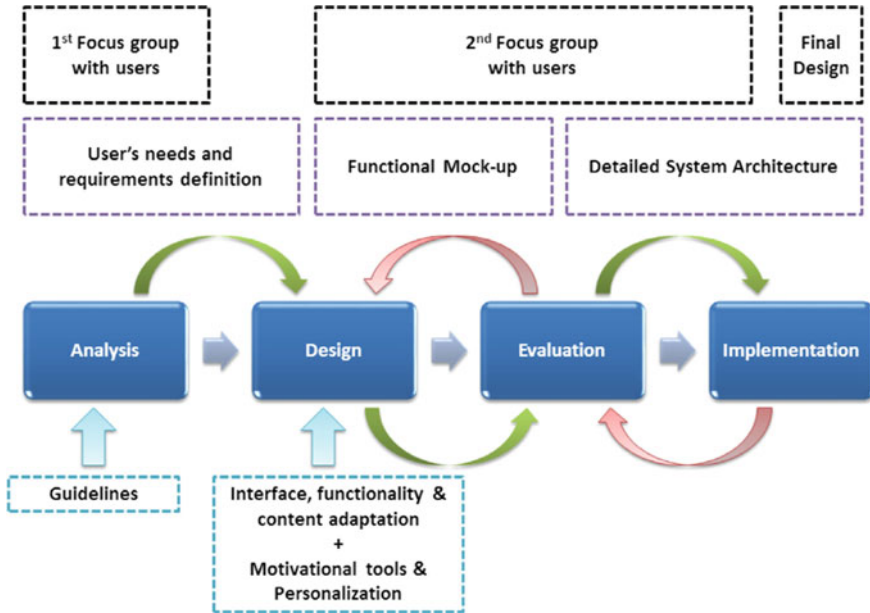


Fig. 9.3 UCD method applied in Smart4MD

In the research stage, referring to the published literature is an excellent way to understand the context in which the development is focused. Scientific articles, patents, even commercial products, are basic sources to accomplish this stage. However, it is also worth-knowing first-hand, so doing interviews or consultations with potential users will help to have clearer the goals in terms of usability of the solution.

There are several methods of approaching users to know their needs and expectations, in terms of their health condition and the use of new technologies, which allow to establish the use cases. These methods can be, for example: focus groups, online or telephone surveys, individual interviews, etc. The selection of the method to use will depend on the type of users, considering age, characteristics of a disease, intellectual and physical capacity, and affinity with new technologies.

On the other hand, during the development phase of the solution, it is advisable to carry out validations with potential users, so that any problem that may affect the final solution can be detected on time. For this validation, there is no single rule. Standards such as DIN EN ISO 9241 and ISO 3407 can be applied as well as other usability tests applying Likert scales. The key is to select the best validation technique that really allows to detect both the strengths and weaknesses of the development and that contributes to improve the final solution.

Table 9.1, we provide a summary of all the recommendations for the different phases of the UCD based on the experiences and lessons learned.

**Table 9.1** Recommendations to apply UCD methodology in ICT solutions

| Phase of UCD methodology          | Recommendations   |
|-----------------------------------|---|
| User needs identification         | <ul style="list-style-type: none"> <li>– Research in all possible sources, academics publications, commercial solutions, previous research projects, patents, etc.</li> <li>– Refer to potential stakeholders, learn from their own experience</li> <li>– Avoid mixing up groups of people to design an all in one product</li> <li>– For dementia groups: reveal technology familiarity issues and special preferences applying UCD at very early stages</li> </ul>                              |
| Technical requirements definition | <ul style="list-style-type: none"> <li>– Always consider the potential users</li> <li>– The solution should be easy to use</li> <li>– If real people are not available for testing, virtual user models deliver promising results and extract requirements and needs, like real users</li> </ul>  |
| Use case and scenarios            | <ul style="list-style-type: none"> <li>– There are no “obvious” use cases, describe them all</li> <li>– Users are key element, trust them. They will give you the best use case possible</li> <li>– Consider all the potential stakeholders; it will be easier to discard one during the development of your solution than create a new one</li> <li>– To produce very detailed and informative scenarios and personas incorporate the literature with EU projects and local knowledge</li> </ul> |
| Develop and validation            | <ul style="list-style-type: none"> <li>– Avoid developing extra functionalities if not asked by the end user</li> <li>– Reuse existing software or adapt it to your needs</li> <li>– There are multiple validations methods, use the one that fits better your goals</li> </ul>   |
| All phases UCD                    | <ul style="list-style-type: none"> <li>– For big-scale pilots, use the co-creation framework and adapt it to your case</li> <li>– To extract IoT proposition, use the co-creation framework</li> <li>– On every iteration on UCD make sure to have solved all the issues discovered before moving to the next iteration</li> </ul>  |

**Acknowledgements** This work has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement ACROSSING No676157.

## References

1. Li Y, Oladimeji P, Monroy C, Cauchi A, Thimbleby H, Furniss D, Vincent C, Blandford A (2011, December) Design of interactive medical devices: feedback and its improvement. In: 2011 IEEE international symposium on IT in medicine and education, vol 2. IEEE, pp 204–208

2. Norman DA, Draper SW (1986) User centered system design: new perspectives on human-computer interaction. CRC Press, Boca Raton
3. Abras C, Maloney-Krichmar D, Preece J (2004) User-centered design. In: Bainbridge W (eds) Encyclopedia of human-computer interaction, vol 37(4). Sage Publications, Thousand Oaks, pp 445–456
4. Ritter FE, Baxter GD, Churchill EF (2014) Foundations for designing user-centered systems. Springer, London, pp 978–981
5. Ghazali M, Ariffin NAM, Omar R (2014, September) User centered design practices in health-care: a systematic review. In: 2014 3rd international conference on user science and engineering (i-USER). IEEE, pp 91–96
6. Teriús-Padrón JG, Kapidis G, Fallmann S, Merdivan E, Hanke S, García-Betances RI, Cabrera-Umpiérrez MF (2018, March) Towards self-management of chronic diseases in smart homes: physical exercise monitoring for chronic obstruction pulmonary disease patients. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, pp 776–781
7. Liappas N, Cabrera-Umpiérrez MF (2018, March) Enabling adaptive interactions to leverage interventions for people with mild cognitive impairments. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, pp 454–455
8. Liappas N, García-Betances RI, Teriús-Padrón JG, Cabrera-Umpiérrez MF (2018, March) Studying the technological barriers and needs of people with dementia: a quantitative study. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, pp 884–889
9. World Health Organization (2018) Noncommunicable diseases [Online]. Available <https://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases>. Accessed 05 Apr 2019
10. Vera-Muñoz C, Arredondo MT, Ottaviano M, Salvi D, Stut W (2013, July) HeartCycle: user interaction and patient education. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 6988–6991
11. Cooper A, Reimann R, Cronin D, Noessel C (2014) About face: the essentials of interaction design. Wiley, Hoboken
12. Fioravanti A (2016) A technological framework based on automatic messaging for improving adherence of diabetic patients. Doctoral dissertation, Telecomunicacion
13. Fico G, Fioravanti A, Arredondo MT, Leuteritz JP, Guillén A, Fernandez D (2011, August) A user centered design approach for patient interfaces to a diabetes IT platform. In: 2011 annual international conference of the IEEE engineering in medicine and biology society. IEEE, pp 1169–1172
14. Vera-Muñoz C, Arredondo MT, Peinado I, Ottaviano M, Páez JM, de Barrionuevo AD (2011, July) Results of the usability and acceptance evaluation of a cardiac rehabilitation system. In: International conference on human-computer interaction. Springer, Berlin, Heidelberg, pp 219–225
15. Antonini A, Gentile G, Giglio M, Marcante A, Gage H, Touray MM, Fotiadis DI, Gatsios D, Konitsiotis S, Timotijevic L, Egan B (2018) Acceptability to patients, carers and clinicians of an mHealth platform for the management of Parkinson’s disease (PD\_Manager): study protocol for a pilot randomised controlled trial. *Trials* 19(1):492
16. Brand CS (2012) Management of retinal vascular diseases: a patient-centric approach. *Eye* 26(S2):S1
17. Scott MJ, Spyridonis F, Ghinea G (2015) Designing for designers: towards the development of accessible ICT products and services using the VERITAS framework. *Comput Stand Interf* 42:113–124
18. García-Betances R, Cabrera-Umpiérrez M, Ottaviano M, Pastorino M, Arredondo M (2016) Parametric cognitive modeling of information and computer technology usage by people with aging-and disability-derived functional impairments. *Sensors* 16(2):266

19. Korn P, Bekiaris E, Gemou M (2009, July) Towards open access accessibility everywhere: the AEGIS concept. In: International conference on universal access in human-computer interaction. Springer, Berlin, Heidelberg, pp 535–543
20. Ignacio Madrid R, Carmona I, Montalvá Colomer JB (2015) Managing the participation of people with disabilities in large-scale R&D technology projects: best practices from AEGIS and CLOUD4ALL. *J Accessib Des All* 5(2):77–99
21. Van Isacker K, Slegers K, Gemou M, Bekiaris E, (2009, July) A UCD approach towards the design, development and assessment of accessible applications in a large scale European integrated project. In: International conference on universal access in human-computer interaction. Springer, Berlin, Heidelberg, pp 184–192
22. Gemou M, Bekiaris E (2009, July) Evaluation framework towards all inclusive mainstream ICT. In: International conference on universal access in human-computer interaction. Springer, Berlin, Heidelberg, pp 480–488
23. Koehler R, Wilhelm E, Shoulson I (eds) (2012) Cognitive rehabilitation therapy for traumatic brain injury: evaluating the evidence. National Academies Press, Washington DC
24. Panou M, Cabrera MF, Bekiaris E, Toulou K (2015) ICT services for prolonging independent living of the elderly with cognitive impairments-IN LIFE concept. *Stud Health Technol Inform* 217:659–663
25. Panou M, Garcia A, Bekiaris E, Toulou K (2016) From user needs and requirements to use cases for ICT services addressed to elderly with cognitive impairments. In: The international eHealth, Telemedicine and Health ICT Forum for Education, Networking and Business—MEDETEL, Luxembourg, 6–8 Apr 2016
26. Medrano-Gil AM, de los Ríos Pérez S, Fico G, Montalvá Colomer JB, Cea Sánchez G, Cabrera-Umpierrez MF, Arredondo Waldmeyer MT (2018) Definition of technological solutions based on the internet of things and smart cities paradigms for active and healthy ageing through Cocreation. *Wireless Communications and Mobile Computing*, 2018
27. Fico G, Montalva JB, Medrano A, Liappas N, Mata-Díaz A, Cea G, Arredondo MT (2017) Co-creating with consumers and stakeholders to understand the benefit of internet of things in smart living environments for ageing well: the approach adopted in the Madrid Deployment Site of the ACTIVAGE Large Scale Pilot. In: *EMBEC & NBC 2017*. Springer, Singapore, pp 1089–1092
28. Frögren J, Quitana M, Anderberg P, Sanmartin Berglund J (2018) Designing a model app for older persons with cognitive impairment: insights from a usability perspective. *Gerontechnology* 17:80



# Chapter 10

## Service Robot Behaviour Adaptation Based on User Mood, Towards Better Personalized Support of MCI Patients at Home



**Dimitrios Giakoumis, Georgia Peleka, Manolis Vasileiadis, Ioannis Kostavelis and Dimitrios Tzovaras**

**Abstract** A novel affective policy has been developed for a service robot, which emphasizes on assistance scenarios focusing on the needs of persons with mild cognitive impairment (MCI) and at early Alzheimer’s disease (AD) stages, at home. This chapter introduces a theoretical framework whose main contribution is twofold; the first one concerns a study on detecting user emotions during human–robot interaction, and the second one concerns the translation of the detected user emotions, into short-term (e.g. half-hour) and long-term (e.g. daily-long) user mood estimates. Alongside, mappings are established between recognized user affect and effects on the robot’s cognitive functions that drive the robot’s assistive behaviour. This results into either the engagement of the service robot in some assistive intervention, aiming to induce more positive outlooks, e.g. stimulating a contact with relatives or friends, or modifications in the way that the robot provides assistance so as to induce a positive feeling of care to the user. Within this chapter, we also propose specific application scenarios of Ambient Assisted Living, on which the current theoretical study can be applied, enabling a service robot to adapt its assistive behaviour based on its user mood and emotions, enabling thus the user to maintain positive affect where possible.

**Keywords** Ambient assisted living · Affect recognition · Robotic assistant · Mild cognitive impairment

### 10.1 Introduction

The world population is ageing and, consequently, the ratio of older persons in modern societies grows, making the capacity to provide proper caregiving services through human caregivers become continuously reduced. Service robots, capable to help older persons in their daily activities at home, could help alleviate this in the

---

D. Giakoumis (✉) · G. Peleka · M. Vasileiadis · I. Kostavelis · D. Tzovaras  
Centre for Research and Technology Hellas, Information Technologies Institute, 6th Km  
Charilaou-Thermi Road, 57001 Thermi-Thessaloniki, Greece  
e-mail: [dgiakoum@iti.gr](mailto:dgiakoum@iti.gr)

future [1]. Such robots could share some responsibilities with the human caregiver of those in need, for some periods of time, supporting a series of activities focusing on AAL. They can provide assistance provision to daily activities such as cooking, eating and medication, through proactive and discrete monitoring of the end-user's activities and robot interventions by reminders and robotic manipulations when deemed necessary [2, 3]. Given that, these robots will also have a role of a companion for the time that they provide their services, it is important for their behaviour to encompass among others a sense of empathy, endorsing affective aspects in the human–robot interaction.

Following the rapid expansion of the field of affective computing [4], the role of affect in human–robot interaction has been an active research topic during the last decade [5]. Typically, research efforts focus on the one hand on the capability of robots to sense emotions of their human counterparts [6, 7], while on the other, on the robots' capacity to express emotions [8, 9], towards emotionally rich human–robot interaction experiences. In the present study, we specifically focus on the challenge of enabling a service robot to draw inference on the mood of its user and on this basis adapt its assistive behaviour, as realized through human–robot interaction framed in the context of AAL-oriented support scenarios, specifically while the robot provides multifaceted, proactive assistance to an older person at home. In particular, we focus on the case of a service robot that aims to assist older persons with MCI and at early stages of AD, persons who typically face, among others, significant memory problems.

According to a recent systematic review with meta-analyses on MCI and mood [10], symptoms of depression and anxiety are considered more prevalent in people with MCI than in people with normal cognitive function and may increase the risk of progression from no cognitive impairment to MCI. Given the significant role that positive emotions play in resilient ageing, early AD patients may retain abilities to achieve subjective well-being in this respect, despite the cognitive decline. In this line, rehabilitation and training programs for AD patients may capitalize on their preserved emotional capacities to reduce or compensate for the cognitive deterioration [11]. On the other hand, the provision of more encouragement and positive responses to AD patients may serve as a complementary and alternative medicine for them, helping them to build up resilient adaptation [12]; in that work, it had been concluded that laughter and smiling associated with pleasant feelings could be of particular benefit.

Along this line, a series of studies have focused on the use of communicative robots, such as the seal-type mental commitment robot, *Paro*, in the support and treatment of dementia patients [13], as well as on cognitively healthy older adults [14], demonstrating that robot-assisted therapy can even have potential on improving the condition of brain activity in patients suffering from dementia [15].

Following the above, some key implications for assistive robots targeting the support of MCI and early AD patients, as well as for the aims of the present study come to light. Overall, empathic communication channels between the human and the assistive robot, endorsing both robot affective input (i.e. recognizing the user's affective state) and output (i.e. affect-related actions of the robot) can be important parameters towards further supporting the user's emotional and overall well-being,

upon the robot's presence. In this scope, enriching human–robot communication with affective reactions (e.g. facial expressions) stemming from the robot can further facilitate the user to reside in an emotionally rich environment, even when s/he is alone with the robot, promoting positive affect where possible. Moreover, detecting user emotions, especially the presence or dominance of the effect of negative valence, is essential towards robotic behaviours that try to counteract and induce more positive outlooks to the user.

Focusing on the latter part, a methodology related to an affective policy is presented in the present chapter. It maps detected user affective states, emphasizing on negative user mood, to robot actions in order to induce more positive emotions to the user, through changes in the robot's behaviour when needed; e.g. from stimulating the user to contact relatives and friends, through to driving more proactive behaviours of the robot so as to induce a more positive feeling of “being cared”, while the service robot provides its assistance interventions.

## 10.2 Related Work

Due to the complexity of emotions [16–18], as well as their highly personalized nature, their automatic recognition through computer and robotic systems is in general a far by trivial task, which has received extensive focus from research, especially during the last decades. As further explained below, in the present work we consider that the service robot includes (possibly multimodal) channels for user affect recognition, thus our developed framework is fed with the input of user emotions detected through such methods.

On the other hand, the affective output that can be provided by service robots can be considered as the simulation and expression of human emotion in robots. This “emotion synthesis” includes several potential channels of expression, such as facial animation, speech, gesture and others. While there is significant past research on robotic gestures as a means of affective output [19], in the present work, we focus more on how the robot actions themselves, in the context of its assistance provision interventions can be affected through the recognized user affect. More specifically, while we consider that a service robot, e.g. like RAMCIP [20], can include affective output interfaces in the form of a facial display, or affect-driven voice changes in speech intonation, we also consider that the assistive behaviour of a service robot can adapt accordingly, following detected changes in the user's affective state.

According to the theory of emotion reappraisal [21–25], the cognitive emotion regulation [26, 27] is applied when a person can regulate emotions by choosing situations, sub-situations, aspects and meanings that may in the direct or indirect way change the emotion gained in a particular occasion. On the ground emotion that has been risen, the cognitive construct moderates the state of the person. As well through the intentional actions, the individual may decrease/increase the intensity of the present emotion and even change its valence. This can lead to the suppression of negative emotions and promotion of more positive moods. For instance, upon the

emergence of negative affective states, such as stress or sadness, a person may choose to get involved in a situation with a specific expected emotional value, so as to change her/his mood level and thoughts [28]. Indicatively, by having an older person living alone, choosing to meet with her/his children or grandchildren, a beloved relative or friend, thus selecting to get involved within a usually positive experience, the mood level can be improved and the senior's thoughts can be positively influenced [25].

Nevertheless, not all persons can sufficiently undertake emotion regulation tasks as the ones described above when needed. It is acknowledged that the extent to which a person can get properly involved within such tasks depends on the person's sensitivity, which can be considered as the ability to change or choose situations in order to bring the person's mood level closer to a prospected positive one [25]. As further explained previously, MCI and AD patients have greater difficulties to deliberately regulate and control their emotional responses than healthy older adults do [8]; the deterioration of cognitive functions can hamper controlled emotional suppression.

In order to help the target audience with enhancing their coping strategies for emotion regulation when needed, one of the target use cases of a service robot can concern the provision of stimulation to the user to contact a relative, upon the detection of prolonged sadness or stress during the day. Alongside, we further propose herein that by maintaining a second, shorter-term index of the user's mood, derived through the emotions monitored during human–robot interaction, the robot can alter its assistive behaviour while providing assistive interventions, in order to induce a more positive feeling of care to the user, trying as such to lead into more positive user emotions.

## 10.3 User Mood Inference Framework Overview

### 10.3.1 *Robot Affective Input and Emotion Recognition Modalities*

During human–robot interaction and monitoring phases, there are several modalities a service robot could exploit in order to assess its user's mood. Facial expression recognition is one of the most practised methods in social robotics [29], apart from gestures and behavioural cues [30]. Herein, the modalities discussed originate from RAMCIP robot's empathic communication channels, comprising of three different mechanisms of affective input.

Specifically, video streams of the user's face, body and biosignals were collected using two different types of sensors. A video camera (Kinect) placed on the robot's head was utilized to get recordings of the user's face and body activity. These data were used to extract inference of the user's mood through facial expression recognition and affect-related body activity analysis. An Empatica E4 wristband placed at the user's hand was employed to record the user's biosignals. This wristband transmits

wirelessly the data of the monitored biosignals to the robots since it is equipped with a series of sensors, such as an electrodermal activity sensor for obtaining galvanic skin response (GSR) and a photoplethysmography sensor for tracking blood volume pulse (BVP). The acquired biosignals, facial expressions and body activity feature recognition are the core modalities of the proposed framework and act as input to a cluster of classifiers in order to get the final inference of the user's current emotional state.

As is further explained below, in the present work, we are particularly interested in the recognition of specific user emotions, of either positive or negative valence, such as joy, sadness, anger and psychological stress. As such, features derived from the user's detected facial expressions, body activity and biosignals, in line with past relevant works [30–33], are fed within our framework, into an SVM-based classification scheme, operating on the basis of multimodal late fusion.

### ***10.3.2 Situations of Interest for Emotion Recognition Towards Mood Estimation***

The **situations of interest** in our case derive from the RAMCIP robot's target use cases which are related to cases where interaction between the user and the robot is established. At this point, it should be noted that affective cues such as facial expressions or affect-related body activity, on which the RAMCIP emotion recognition (ER) methods are based, are typically event-triggered, communicative or conversational signals [34, 35].

Moreover, it should be noted that the recognition of the target emotional states of joy, sadness and anger relies strongly on the facial expressions and body activity modalities. As such, it becomes clear that in order to anticipate more reliable recognition of these affective states in the target realistic settings of RAMCIP, the situations of interest should focus on human–robot communication cases. In addition, on cases where the user is engaged at a task that allows for a clear view of the user's body and face should be considered; such as when the user is for instance playing a cognitive exercise game that is provided by the RAMCIP robot, namely the “Virtual Supermarket Test” [36].

Following the above, the recognition of the target emotional states is attempted in the situations of interest, which can be grouped in the following main categories:

1. HRI: the robot is engaged in an HRI task, where communication with the user is also involved; e.g. from the provision of a notification to the user or in more general, establishment of a human–robot communication scenario, where non-verbal user affective cues may appear during HRI.
2. Games: the user is playing a cognitive exercise game, which is provided by the robot.

Alongside, given that stress recognition is mainly based on the monitoring of the user's biosignals, this affective state falls in a different category, where affect

recognition can be attempted also in further cases, where the user is not in direct, visual contact with the robot. Therefore, while psychological stress is also monitored during the HRI and cognitive games playing cases, it can as well be monitored during cases in the daily life of the user at home.

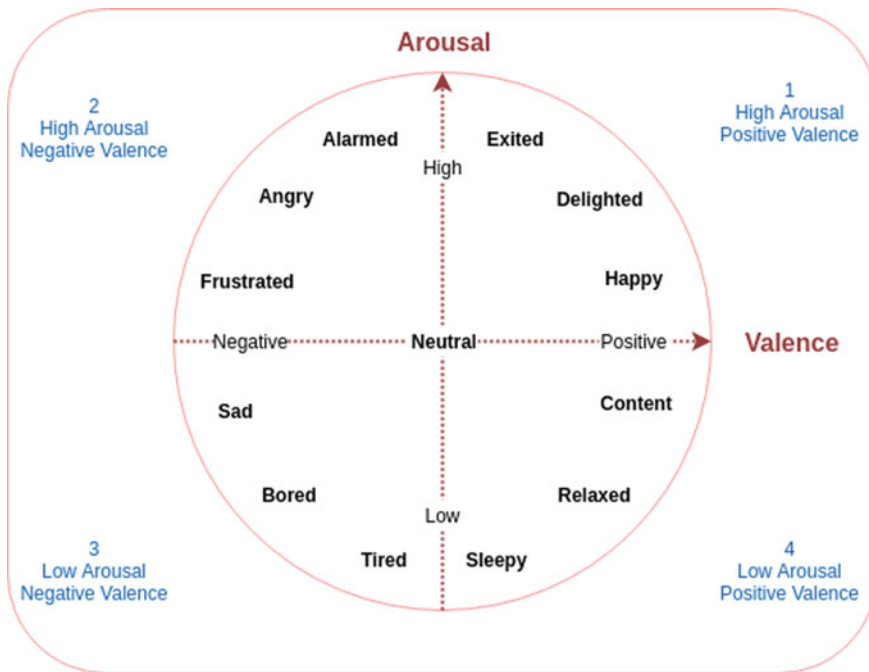
### 10.3.3 From Detected User Emotions to User Mood Inference

The detected user emotions, as recognized by the emotion recognition modalities, have a threefold purpose in the overall robot's operation; specifically, the focus is on the detection of affective states either related to psychological stress or indicative to user discomfort, with the aim to augment the robot cognitive functions towards assisting the user. The inferred emotions also provide the robot with feedback over the effect that action had on the user, populating with this information the user's model, which the robot uses to adapt its behaviour in a personalized way.

In the formulation of our robot affective policy, the detected user emotions are also combined to lead into estimates on the user's overall **mood**, both in the **short-term** (i.e. referring to the "last hour's" time period) and in the **long-term** (i.e. referring to a single day). The outcomes of both inference processes are used to augment the cognitive functions of the robot towards assisting the user.

Specifically, the outcome of the **long-term** (daily-based) inference focuses on the augmentation of the robot's cognitive functions with the capability to decide on whether stimulation of the user to contact a relative should be provided, so as to help the user with her/his coping strategies for emotion regulation. Moreover, the outcome of the **short-term** (e.g. hourly based) inference on the user's mood is used so as to drive alterations on the robot's assistive behaviour, in essence, the short-term inference is used to help the robot modify the way that assistance is provided to the user, so as to demonstrate a more proactive and helpful behaviour through its current task, towards inducing a more pleasant feeling of care in case that negative user emotional states are detected.

Overall, our mood estimation model is based on the valence–arousal (VA) space [37], where valence and arousal are considered to span a Euclidean space. We consider that the target emotional states of happiness, sadness, anger and stress are mapped to specific quadrants of the VA space (Fig. 10.1). While happiness belongs to the *quadrants of positive pleasure*, sadness and anger belong to the *negative ones*. As concerns the targeted affective state of psychological stress, it can be considered in general to arise when the person realizes that the environmental demands surpass her/his adaptive capacity to meet them. Therefore, our approach is based on the transformation of the detected target emotional states into an indication of whether the person's overall mood is of positive or negative valence for the time period of interest.



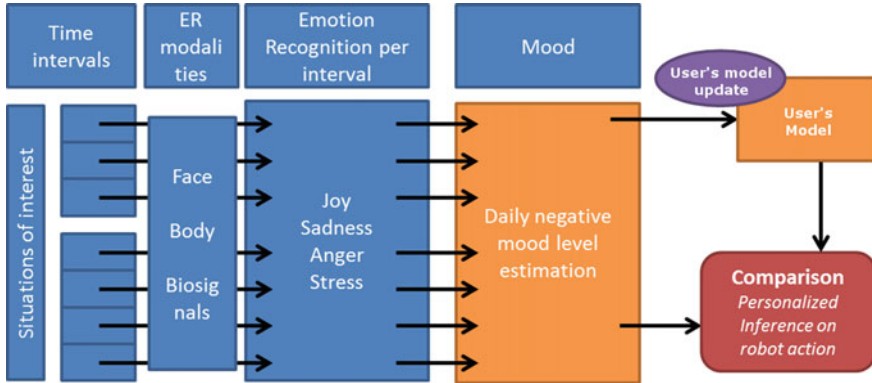
**Fig. 10.1** Affective states within Russell’s circumplex model of affect [10]; horizontal axis: valence, vertical axis: arousal

## 10.4 Affect-Driven Inference on Robot Behaviour Adaptations

### 10.4.1 Long-Term User Mood Inference

In the scope of long-term user mood inference, we have developed an inference model that drives assistive robot decisions, as illustrated in Fig. 10.2 and further explained in what follows.

The emotion recognition results that derive from the processing of the different modalities (face, body and biosignals) for specific time intervals of interest are provided as input to the “daily negative mood level estimation” module. Therein, through the mood estimation model described below, the individual results are merged so as to formulate an index of the user’s negative mood level for the current day. Subsequently, this resulting estimate is compared to the user’s typical daily negative mood level index. Upon the detection of a significant difference, the robot’s corresponding actions to support the user in improving her/his mood are triggered.



**Fig. 10.2** Overview of the personalized inference model for user stimulation to contact a relative so as to improve her/his mood

### 10.4.1.1 Long-Term Negative Mood Inference Model

For the purpose of long-term mood inference, we follow the rationale of the “Long” (longer emotion) mood predictor defined in [38], which is based on the hypothesis that the quadrant of the PAD space [39] containing the majority of detected emotions may be a predictor of the user’s mood.

The core factor of our **long-term** mood inference model is the *NvA* (Negative versus All) index (10.1), which is calculated as the ratio between the occurrences of detected user emotional states of negative valence, i.e. those that belong to the VA quadrants of *negative* valence values (negative), to the total count of intervals where emotion recognition was attempted to take place:

$$NvA = \frac{\text{Count}(e(V < 0))}{\text{Count}(\text{ALL})} \tag{10.1}$$

where  $e(V < 0)$  corresponds to the detected emotional states: “Sad”, “Angry”, “Stressed”, while “ALL” comprises all intervals of interest where emotion recognition was attempted.

As described above, there are two different types of situations of interest that are involved in our ER approach, the “HRI” and “Games” situations. Moreover, in the case of RAMCIP robot, there is a key communication between the robot and the user in the beginning of the proactive scenario of the “Communication with relatives and friends” use case, where the robot asks “How do you feel today?”, in case that some signs of inactivity and/or negative mood have been detected. Following the above, we define the following indexes:  $NvA_{HRI}$ ,  $NvA_{Games}$  and  $NvA_Q$ , which correspond to the calculation of the *NvA* index for the total time intervals of the HRI, the cognitive games and the question, respectively. The total user’s *NvA* index (10.2) for the current day ( $D$ ) is then calculated by applying a weighted average on the above three factors, i.e.



$$NvA(D) = w_1 \cdot NvA_{HRI}(D) + w_2 \cdot NvA_{Games}(D) + w_3 \cdot NvA_Q(D) \quad (10.2)$$

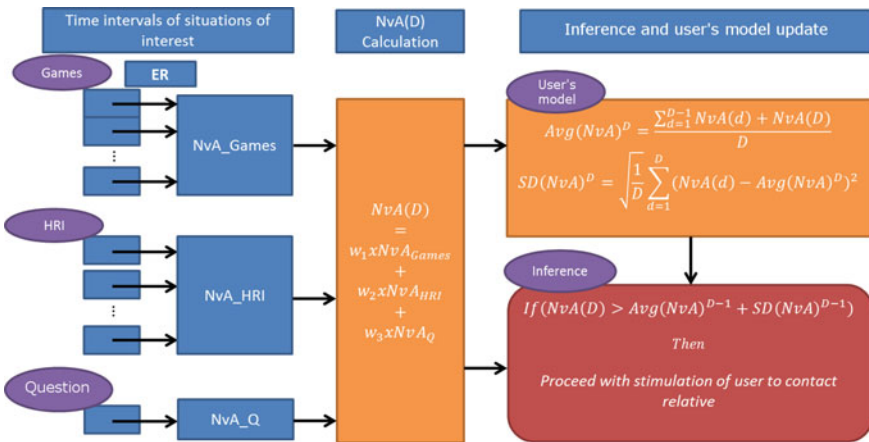
Clearly, as each of the weights  $w_1$ ,  $w_2$  and  $w_3$  increases, the impact of the user’s detected emotions within either, the HRI sessions, the cognitive games, or the question, increases, respectively. In our current formulation, we have defined  $w_1 = 0.3$ ,  $w_2 = 0.3$ ,  $w_3 = 0.4$ .

Through the above set-up, there is an equal contribution between the user’s detected emotions during the HRI sessions and the cognitive games, while a stronger contribution is provided from the user’s detected emotional state while answering the question. This comes in analogy to the rationale of the “complex mood predictors” defined in [30], where the mood assessment gradually neglects the past and gives more weight to the most recent detected affective states.

**10.4.1.2 Personalized Inference Towards Robot Affect-Related Actions, Relevant to Stimulation of User’s Communication**

As shown in Fig. 10.3, the history of past  $NvA$  index values for the past days of the robot’s operation (i.e. for the  $D - 1$  days, assuming that  $D$  is the current day) is recorded in the user’s model, while their average  $(Avg(NvA)^{D-1})$  and standard deviation  $(SD(NvA)^{D-1})$  are the parameters that are used in the robot’s inference process for actions stimulating the user to contact a relative.

Specifically, upon the calculation of the user’s  $NvA$  index for the current day ( $NvA(D)$ ), a comparison takes place to the typical corresponding user’s behaviour encoded in their model; it is checked whether the user’s current  $NvA$  index ( $NvA(D)$ ) has a significant difference (i.e. exceeding the standard deviation) to the average user’s  $NvA$  values known so far. In case where the user’s current mood is found to



**Fig. 10.3** Functional overview of the personalized inference model for user stimulation to contact a relative so as to improve her/his mood

involve negative emotional states to the extent that bring it to a significant difference than her/his average daily mood, the affect-related robot actions are triggered, leading the robot to stimulate the user to contact a relative or friend.

Following the above inference process, an update of the user's current  $\text{Avg}(NvA)$  and  $\text{SD}(NvA)$  values (as had been calculated from the days 1, 2, ...,  $D - 1$ ) takes place, so as to take into account in the next iteration also the new values calculated for the user. This way, the inference model for the triggering of the robot's corresponding action is adapted to the detected user's affective behaviour, as it evolves over time.

### 10.4.2 Short-Term Mood Inference

While the above, long-term mood inference model is used to provide estimates on the overall user's mood on a daily basis, a further model has also been developed dedicated to provide continuous, shorter-term (hourly based) estimates on the user's mood as it evolves over time during the day. This derives from an effort to further augment the robot's cognitive functions with affect-related knowledge derived from user monitoring.

The overall rationale of this process is based on the assumption that when the user's current mood is detected to be at a negative level, the robot's assistive behaviour would be of great importance to become more proactive and prone to establish interventions that involve intervention with robotic manipulations rather than with communication in an effort to induce her/him with a more pleasant feeling of care [40], similarly to the alterations that are made in the robot's assistive behaviour upon the detection of user physical fatigue.

While the long-term mood estimation model described above is based on the  $NvA$  index, the short-term mood inference case has a basic difference, as our interest is on the more direct estimation of the short-term (i.e. on an hourly basis) user's mood, towards altering the way that assistance is provided by the robot. Specifically, in this case, our model is based on the difference between the occurrences of positive and negative affective states, having thus the " $PvN$ " (Positive versus Negative) index described below as its core factor, instead of the  $NvA$  index that is used in the long-term model. Moreover, an exponential discount function is used so as to weight within the  $PvN$  index calculation the different ER outcomes that have derived for the period of interest, giving more weight to the most recent ones, as in this short-term case, no explicit communication is conducted between the robot and the user so as for the robot to be provided with a more definite, current indication from the user on her/his current mood.

By using this  $PvN$  index, we are not only capable of estimating the short-term mood of the user on a positive versus negative basis, but we can use this same index so as to encode the affective feedback derived from human-robot communication during specific states of the target use cases, denoting whether specific actions of the robot typically have either positive or negative affective effect to the user. By combining the knowledge on both the user's current state in terms of positive versus

negative mood, with the known effects of robot actions, the robot is then capable to infer on its assistive behaviour in a way that tries to avoid negative emotions to emerge to the user and further help her/him have more positive outlooks.

### 10.4.2.1 Short-Term Mood Inference Model

For the purpose of short-term mood inference, we follow the rationale of the “exponential discount” (ED) mood predictor defined in [30], which is based on the hypothesis that a desirable property of mood assessment is smoothness over time [41]. This has been modelled in the ALMA model [42] through a linear discount function; however, the results of [30] indicated that the use of exponential discount can lead to even better results on mood estimation. Following the above, our short-term mood inference model formulation is as follows:

Let  $E_h = e(1), e(2), \dots, e(i)$  be the time series of the user’s target emotions (i.e. joy, sadness, anger, stress), detected in the  $i$  short (1-min long) time intervals that belong to the last hour of observations. By considering whether each detected emotion belongs in the positive or the negative half of the valence axis in the VA space, we replace each element of  $E_h$  with the value of “+1” or “−1”, respectively, and obtain the following time series:

$$V_h = v(1), v(2), \dots, v(i) \quad (10.3)$$

where  $v(i)$  is either  $-1$  or  $+1$ , if the corresponding detected emotion detected within the time interval  $i$  of the last hour ( $h$ ), belongs to the negative or the positive half of the VA space, respectively (10.3).

We define the exponential discount function  $D_w$  (10.4) in our case as:

$$D_w(k) = \exp\left(\frac{k-1}{i-1}\right), k = 1, 2, \dots, i \quad (10.4)$$

Our short-term mood index (**P**vs**N**( $i$ )—**P**ositive versus **N**egative index (10.5), for the last  $i$  emotions) is then calculated as the weighted average of the contents of the  $V(i)$  time series:

$$PvN(i) = \sum_{k=1}^i v(k). \quad (10.5)$$

By considering the last  $i$  emotions detected during the last hour ( $h$ ), we end up with the  $PvN_h$  index (10.6), which is formally defined as:

$$PvN_h = PvN(i) \quad (10.6)$$

where  $i$  denotes the user emotions recognized during the last hour of the robot's operation.

If the  $PvN_h$  index is found to be negative, this leads in our case to a negative short-term mood estimate for the user, with respect to the corresponding time period. In this case, a corresponding signal may be provided to the robot's cognition module, signifying this fact and leading the robot cognitive functions to bias the robot's behaviour towards a more proactive one, more prone to provide assistance through robotic manipulations, than through communication.

#### 10.4.2.2 Encoding and Using Affective User Feedback on Robot Actions

By considering the target use cases of a service robot like RAMCIP [17], a series of occasions can be identified where the robot's proactive behaviour can be prone to induce either positive or negative affect to the user. More specifically, we take into account that: (a) the robot's proactive behaviour can either focus on providing physical assistance to the user (e.g. to bring some object to the user or change the state of some electric appliance through robotic manipulations), or on stimulating the user to perform the corresponding action on her/his own, so as to be more physically active, in cases where no user fatigue is detected. (b) On the other hand, having the robot demonstrating constantly a behaviour that involves insisting on stimulating the user to act may lead at some point to the emergence of negative emotions to the user.

It thus becomes clear that the robot cognitive functions should maintain a balance between the assistance that is provided through human-robot communication and the more proactive application of robotic manipulations, within the target use cases where assistance can be provided in both of these ways. Apart from the user's physical fatigue, which is clearly a key factor in this respect, further important factors, from the psychological point of view, can be considered both (a) the user's current mood and (b) the known associations of robot actions to the user's affect.

Following the above, our inference model concerning the effects of the user's affective cues on the short-term robot decisions over how assistance is provided within the target situations of interest relies on two basic factors, i.e. the  $PvN_h$  index described above and a further PvN-based factor, calculated through the detected user emotions upon specific robot actions of interest. More specifically, we consider the following two factors:

- $PvN_h$ : estimate on the user's current mood, in respect to positive versus negative emotions detected in the short-term (i.e. during the last hour)
- $PvN(CoI(A), Comm(C)) = PvN_{A,C}$ : PvN index calculated through the detected emotions upon past appearances of the human-robot communication scenario  $C$ , involved in the case of interest (CoI)  $A$ .

Let  $e(CoI(A), Comm(C))^i$  be the user emotion recognized during the  $i$ th appearance of the human-robot communication state of interest  $C$  in the target  $CoI A$ . The  $PvN_{A,C}$  index is calculated equivalently to the  $PvN(i)$  (as described in the previous

section), by using in this case the time series:  $E_{A,C} = e(A, C)^1, e(A, C)^2, \dots, e(A, C)^i$ , instead of  $E_h$ .

Following the above, the robot stores for each *CoI* and communication of interest the value  $E_{A,C}$ , which indicates the weighted average of the user's emotional responses. Each  $E_{A,C}$  parameter is given 0 as the initial value, i.e.  $e(A, C)^0 = 0$ .

The overall rationale of the use of the above parameters towards affecting the robot's cognitive functions is as follows:

In case that (a) the user's current mood (as expressed through  $PvN_h$ ) is detected to be of negative valence (i.e.  $PvN_h < 0$ ) AND a robot's behaviour, stimulating the user to act through communication instead of involving robotic manipulations, so as to resolve a specific situation within a target situation of interest, has been detected in previous occurrences of this situation to be associated with negative user affective cues (e.g. anger, stress, sadness detected upon the specific human-robot communication), thus being associated with a negative  $PvN_{A,C}$  index ( $PvN_{A,C} < 0$ ), THEN, the robot's cognitive functions will promote the selection of a robotic manipulation-oriented assistive behaviour in the given case.

The latter is realized through the provision of a signal to the robot's cognition module, signifying that a robotic manipulation-oriented assistive behaviour should be preferred in the current case.

In parallel to the above, the update of the  $E_{A,C}$  time series related to the current use case incidence will occur, on the basis of the assumption that emotion of zero (neutral) valence should be added at the point of " $i + 1$ ". This assumption is made due to the fact that the specific communication of interest is avoided and not conducted. In some cases, the above, discarded communication may be replaced by an indication of the robot that will conduct the necessary action on its own so as to help the user; although it could be considered that the (most probably) positive affective outcome detected during the latter communication may be taken into account instead, in that case, the stimuli would have been different, and as such, a further uncertainty would be added on the relevance of the specific ER procedure to the calculation of the specific  $PvN_{A,C}$  index.

## **10.5 Application of Mood Inference Models by Adaptation of Service Robot Use Case Scenarios Based on Mood Inference**

### ***10.5.1 Long-Term Mood Inference During Day 1***

An important issue in the formulation of the long-term mood index concerns the way that the decision on the robot action is taken the first day of the robot's operation with a specific user, i.e. when  $D = 1$ , as during that day, no prior  $NvA$  values exist at the user's virtual model. In order to set this initial threshold, we take into account that, following the study of Carstensen et al. [43], the frequency of negative affect during daily life in persons between the age of 65–94 years old can be considered to reside at the level of 25%. Therefore, the initial threshold, to which  $NvA(1)$  is compared so as for the robot to take a decision on day 1 of its operation is set to 0.25, while each day this initial assumption will gradually be fitted to the observations derived through the robot's emotion recognition engine, over the detected emotions of its actual, specific user.

Notably, the above supposition bears a resemblance to the behaviour of a human caregiver, who, during the early days of providing care to a patient, depends on her/his own experience, comparing the patient's behaviour with the one of others already known, while later the caregiver's decisions become gradually better adapted to the specific user.

### ***10.5.2 Stimulating the User to Contact a Relative Based on Long-Term Mood Inference***

In the case of RAMCIP, the robot stimulated the user to contact a relative in two different ways, either on a scheduled basis at a specific time point during the day or proactively, at time points that are deemed necessary. In the first case (scheduled), the robot is scheduled to stimulate the user in the evening, in case that the user is detected that s/he has not made any call through the robot to a relative or friend during the day. In the second case (proactive), the robot monitors first of all whether psychological stress has strongly appeared to the user within a time period. Specifically, the scenario is triggered in case that a significant proportion of the user's past stress-related measurements reveal increased psychological stress.

### 10.5.3 *Adaptation of Robot Actions Based on Short-Term Mood Inference*

In line with the above formulation, the specific affect-related feedback that is obtained by the robot influences specific robot actions, involved in specific situations of interest. The robot can assess the affective outcome that its actions have on the user, and this information populates the user's virtual model.

Identifying all cases where: either a positive or negative emotion could be detected from the user following a specific proactive assistive robot action where either human-robot communication or robotic manipulations could be involved, and the corresponding derived knowledge would be meaningful to have an effect on corresponding future robot actions, so as make them more prone to induce pleasant emotions to the user when needed.

### 10.5.4 *Simulated Inference Examples on Situation of Interest*

In what follows, two examples are provided, showing how the long-term and short-term mood inference processes can be applied in practice. In both cases, the user is assumed to have the same, significant amount of negative detected emotions during the first day of the robot's operation (i.e. the threshold to which  $NvA(D)$  is compared is 0.25).

#### **Case 1: Long-Term Mood Inference—Contacting a Relative**

$NvA_{Games}(D) = 0.5$ ; i.e. negative emotional states (e.g. stress, anger) have been detected in half of the time intervals involved in the cognitive exercise games played during that day.

$NvA_{HRI}(D) = 0.3$ ; i.e. in one out of the three HRI scenarios established today by the robot, a negative emotional state of the user (e.g. anger) was detected.

$NvA_Q(D) = 1$ ; i.e. during the communication posed to the user, sadness was detected.

Using  $w_1$ ,  $w_2$  and  $w_3$  as defined above, the above values of the  $NvA_{Games}$ ,  $NvA_{HRI}$  and  $NvA_Q$  factors result to  $NvA(D) = 0.64$ . This result is significantly above the initial threshold of 0.25, and thus, the robot decides to stimulate the user to contact a relative.

#### **Case 2: Short-Term Mood Inference—Adapting Robot Actions**

The point of interest in this case concerns the state where the robot asks the user if s/he would like it to undertake a robotic manipulation on behalf of the user. In case that the user has been detected as not tired, the robot shall stimulate her/him into the corresponding action. However, in case that the user's current state is associated with a negative mood, and this specific communication has been found to associate with a negative user's emotional response (e.g. anger, stress, sadness), it should better be

avoided for the current instance and be replaced with a more proactive, physically assistive intervention (10.7).

The specific rule for the above is defined as:

$$\text{If } (PvN_h < 0) \wedge (PvN_{\text{case-instance}} < 0) \quad (10.7)$$

The above rule is evaluated by the affective module each time that the situation of interest of this type comes into play, while in case it is found TRUE, the affective module notifies the cognition module that a robotic manipulation-oriented assistive behaviour should be preferred.

## 10.6 Conclusions

The main focus of this work concerned the research and development of a theoretical framework with its respective models for assessing the overall user's mood and introducing corresponding effects to the robot's cognitive functions, in an effort to further help the end-user maintain positive outlooks. In this respect, computational models drawing personalized inference on whether some specific robot behaviours should be triggered in the scope of the target cases of interest were developed. In our framework, both long-term (i.e. on daily basis) and short-term (i.e. on hourly basis) inference on the user's mood is performed, based on the outcomes of the developed ER methods. Moreover, user affect recognition is being attempted at specific human-robot communication parts in specific situations of interest, allowing the robot to gain knowledge on possible affect-related effects of its actions to the user.

All the above inference outcomes are stored in the affective part of the user's virtual model, which encodes corresponding personalized user specificities. By comparing this user-centric knowledge to the current user state (e.g. typical prevalence of negative emotions during a day versus prevalence of user negative emotions in the current day), the proposed affective module can provide significant affect-related cues to the robot's cognition module, so as to either trigger a specific affect-related use case, or modify the way that assistance is provided to the user. Especially for the latter part, the efforts of this study have identified a series of human-robot communication occasions within the robot's target cases of interest, where the robot's behaviour can be changed into a more proactive one, inducing a more pleasant feeling of care to the user.

**Acknowledgements** This work has been supported by the EU Horizon 2020 funded project "Robotic Assistant for MCI Patients at home (RAMCIP)" under the grant agreement no. 643433.



## References

1. Kostavelis I, Giakoumis D, Malasiotis S, Tzovaras D (2015) Ramcip: towards a robotic assistant to support elderly with mild cognitive impairments at home. In: International symposium on pervasive computing paradigms for mental health. Springer, Cham, pp 186–195
2. Kostavelis I, Vasileiadis M, Skartados E, Kargakos A, Giakoumis D, Bouganis CS, Tzovaras D (2019) Understanding of human behavior with a robotic agent through daily activity analysis. *Int J Soc Robot* 1–26
3. Peleka G, Kargakos A, Skartados E, Kostavelis I, Giakoumis D, Sarantopoulos I, Ruffaldi E (2018) RAMCIP-A service robot for MCI patients at home. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 1–9
4. Picard RW (1995) *Affective computing* (No. 321)
5. Beer JM, Liles KR, Wu X, Pakala S (2017) Affective human–robot interaction. In: *Emotions and affect in human factors and human-computer interaction*. Academic Press, pp 359–381
6. Jerritta S, Murugappan M, Nagarajan R, Wan K (2011) Physiological signals based human emotion recognition: a review. In: 2011 IEEE 7th international colloquium on signal processing and its applications. IEEE, pp 410–415
7. Rani P, Liu C, Sarkar N, Vanman E (2006) An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Anal Appl* 9(1):58–69
8. Saerbeck M, Bartneck C (2010) Perception of affect elicited by robot motion. In: *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, pp. 53–60
9. Lee HS, Park JW, Chung MJ (2007) A linear affect–expression space model and control points for mascot-type facial robots. *IEEE Trans Rob* 23(5):863–873
10. Yates JA, Clare L, Woods RT (2013) Mild cognitive impairment and mood: a systematic review. *Rev Clin Gerontol* 23(4):317–356
11. Zhang F, Ho YW, Fung HH (2015) Learning from normal aging: preserved emotional functioning facilitates adaptation among early Alzheimer’s disease patients. *Aging Dis* 6(3):208
12. Takeda M, Hashimoto R, Kudo T, Okochi M, Tagami S, Morihara T, Tanaka T (2010) Laughter and humor as complementary and alternative medicines for dementia patients. *BMC Complement Altern Med* 10(1):28
13. Shibata T, Wada K (2011) Robot therapy: a new approach for mental healthcare of the elderly—a mini-review. *Gerontology* 57(4):378–386
14. Wada K, Shibata T, Saito T, Sakamoto K, Tanie K (2005) Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In: *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, pp 2785–2790
15. Wada K, Shibata T, Musha T, Kimura S (2008) Robot therapy for elders affected by dementia. *IEEE Eng Med Biol Mag* 27(4):53–60
16. Damasio AR (1994) *Descartes’ error: emotion, reason, and the human brain*. Quill. Google Scholar OpenURL Yorkville University, New York
17. LeDoux JE, Damasio AR (2013) Emotions and feelings. In: Kandel ER, Schwartz JH, Jessell TM, Siegelbaum SA, Hudspeth AJ (eds) *Principles of neural science*, 5th edn. McGraw Hill, New York, pp 1079–1094
18. Begley S, Davidson R (2012) *The emotional life of your brain: how its unique patterns affect the way you think, feel, and live—and how you can change them*. Hachette UK
19. Salem M, Kopp S, Wachsmuth I, Rohlfing K, Joublin F (2012) Generation and evaluation of communicative robot gesture. *Int J Social Robot* 4(2):201–217
20. Kostavelis I, Giakoumis D, Peleka G, Kargakos A, Skartados E, Vasileiadis M, Tzovaras D (2018) RAMCIP robot: a personal robotic assistant; demonstration of a complete framework. In: *Proceedings of the European conference on computer vision (ECCV)* pp 0–0
21. Damasio A, Descartes R (1994) *Error: emotion, reason, and the human brain*. Avon, New York, pp 350–412

22. Ochsner KN, Ray RD, Cooper JC, Robertson ER, Chopra S, Gabrieli JD, Gross JJ (2004) For better or for worse: neural systems supporting the cognitive down-and up-regulation of negative emotion. *Neuroimage* 23(2):483–499
23. Suri G, Gross JJ (2012) Emotion regulation and successful aging. *Trends Cognit Sci* 16(8):409–410
24. Tottenham N, Hare TA, Casey BJ (2011) Behavioral assessment of emotion discrimination, emotion regulation, and cognitive control in childhood, adolescence, and adulthood. *Front Psychol* 2:39
25. Wild B, Erb M, Bartels M (2001) Are emotions contagious? Evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. *Psychiatry Res* 102(2):109–124
26. Gross JJ (1998) The emerging field of emotion regulation: an integrative review. *Rev Gen Psychol* 2(3):271–299
27. Gross JJ (2001) Emotion regulation in adulthood: timing is everything. *Curr Dir Psychol Sci* 10(6):214–219
28. Both F, Hoogendoorn M, Klein MC, Treur J (2008) Modeling the dynamics of mood and depression. In: *ECAI*, pp 266–270
29. Chen H, Gu Y, Wang F, Sheng W (2018) Facial expression recognition and positive emotion incentive system for human-robot interaction. In: 2018 13th world congress on intelligent control and automation (WCICA). *IEEE*, pp 407–412
30. Giakoumis D, Drosou A, Cipresso P, Tzovaras D, Hassapis G, Gaggioli A, Riva G (2012) Using activity-related behavioural features towards more effective automatic stress detection. *PLoS ONE* 7(9):e43571
31. Giakoumis D, Tzovaras D, Moustakas K, Hassapis G (2011) Automatic recognition of boredom in video games using novel biosignal moment-based features. *IEEE Trans Affect Comput* 2(3):119–133
32. Giakoumis D, Tzovaras D, Hassapis G (2013) Subject-dependent biosignal features for increased accuracy in psychological stress detection. *Int J Hum Comput Stud* 71(4):425–439
33. Giakoumis D, Vogianou A, Kosunen I, Moustakas K, Tzovaras D, Hassapis G (2010) Identifying psychophysiological correlates of boredom and negative mood induced during HCI. In: *B-interface*, pp 3–12
34. Ekman P (1997) Should we call it expression or communication? *Innov: The Eur J Soc Sci Res*, 10(4):333–344
35. Calvo RA, D’Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans affect comput* 1(1):18–37
36. Zygouris S, Giakoumis D, Votis K, Doumpoulakis S, Ntovas K, Segkouli S, Tsolaki M (2015) Can a virtual reality cognitive training application fulfill a dual role? Using the virtual supermarket cognitive training application as a screening tool for mild cognitive impairment. *J Alzheimers Dis* 44(4):1333–1347
37. Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39(6):1161
38. Katsimerou C, Heynderickx I, Redi JA (2015) Predicting mood from punctual emotion annotations on videos. *IEEE Trans Affect Comput* 6(2):179–192
39. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr Psychol* 14(4):261–292
40. Mitra A, Dey B (2013) Therapeutic interventions in Alzheimer disease. In: *Neurodegenerative diseases*. IntechOpen
41. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. *IEEE Trans Multimed* 7(1):143–154
42. Gebhard P (2005) ALMA: a layered model of affect. In: *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. *ACM*, pp 29–36
43. Carstensen LL, Pasupathi M, Mayr U, Nesselroade JR (2000) Emotional experience in everyday life across the adult life span. *J Pers Soc Psychol* 79(4):644

# Chapter 11

## Using Semantic Markup to Boost Context Awareness for Assistive Systems



Claudia Steinberger and Judith Michael

**Abstract** Considerable effort to manually configure the user's context and too coarse-grained activity recognition results often make it difficult to set up and run an assistive system. This chapter is the result of our experience with the Human Behavior Monitoring and Support (HBMS) assistive system, which monitors user's activities of daily life and supports the user in carrying out these activities based on his own behavior model. To achieve the required context awareness, we join assistive systems with the semantic web to (1) simplify the construction of a personalized context model and to (2) improve the system's activity recognition capabilities. We show how to semantically describe devices and web applications including their functionalities and user instructions and how to represent these descriptions in the web. The advantages of this semantic markup approach for the application of HBMS-System and beyond are discussed. Moreover, we show how personalized and adaptive HBMS user clients and the power of the context model of HBMS-System can be used to bridge an existing activity recognition gap.

**Keywords** Semantic manual · Behavioral assistance · Assistive system setup · Application markup · Context model

### 11.1 Motivation

Assistive systems that monitor user's behavior and support them in performing activities based on behavioral patterns can be very useful in helping people who suffer from any kind of cognitive decline. However, adequate context awareness and a smooth adaptation of an assistive system to the user's environment are prerequisites for its successful installation.

---

C. Steinberger (✉)  
Department of Applied Informatics, Universität Klagenfurt, Klagenfurt, Austria  
e-mail: [claudia.steinberger@aau.at](mailto:claudia.steinberger@aau.at)

J. Michael  
Software Engineering, RWTH Aachen University, Aachen, Germany  
e-mail: [michael@se-rwth.de](mailto:michael@se-rwth.de)

The demands for context awareness are high when it comes to behavioral assistance. These assistive systems can be seen as cyber-physical systems [13]. Every system instance owns its own physical user environment, in which the actions and environmental conditions of the user have to be monitored and supported based on a context model. To record data on situations in the physical world, human activity recognition (HAR) systems are used applying location-, wearable- or object-based sensors or video observation [36]. Assistive systems are not always modular and interoperable with already available open HAR-Systems. They often implement their own HAR-System because the capabilities of existing open ones do not meet their activity recognition requirements.

In any case, the system setup requires adjustment effort for every user in practice, so that the assistive system can achieve the necessary context awareness. Video observation is strongly rejected by users and smart devices with built-in sensors are often not (yet) available in and for households. However, assistive system research projects hardly report on the effort and problems involved in adapting an assistive system and its HAR-Systems to an end user's environment during system setup.

Over the last ten years, the authors have been working on an assistive system project called HBMS (Human Behavior Monitoring and Support) [28]. HBMS-System aims to actively support people in their daily activities including also the use of their preferred web applications. HBMS-System is interoperable with different open HAR-Systems but uses an own internal knowledge base called Human Cognitive Model (HCM), to define, integrate, and store the user's context model [36]. This contribution is the result of the experiences gained in setting up HBMS-System for user's physical environment, observing and supporting user's behavior and shows our solutions to eliminate encountered weaknesses.

An evaluation of HBMS-System revealed in particular the following weaknesses: (1) In order to provide assistance, information about the user's resources in the physical environment like devices or web applications, his or her social and personal situation, and the spatial environment had to be analyzed and mapped to the HCM, when setting up the system. Especially for modeling the devices with their functionalities and user instructions as well as modeling the functionalities and interaction possibilities with web applications, considerable effort was required. (2) The activity recognition results of the used HAR-Systems concerning fine-grained user interactions with devices or web applications were not satisfying for the intended behavioral assistance [40].

In the following, we show the gaps and challenges in the area of context perception in detail and present an approach to overcome these challenges joining assistive systems with the semantic web to simplify the construction of user's context model.

This chapter is structured as follows: Sect. 11.2 presents the context requirements of assistive systems that aim to provide behavior-based support. Section 11.3 introduces the HBMS-System and HCM as well as our evaluation experiences in more detail. It works out two research questions: *RQ1: How is it possible to support the setup of personalized context models for assistive systems in order to keep the manual effort as low as possible?* *RQ2: How is it possible to improve the fine-grained activity recognition capability of assistive systems to be able to support the user in*

*multiple situations?* Section 11.4 deals with related work to answer these research questions. Section 11.5 presents answers to RQ1 and shows how to semantically markup devices and web applications and deals with related advantages. Section 11.6 presents answers to RQ2 and shows how the results of Sect. 11.5 can be used for personalized and adaptive user clients to boost activity recognition results. Section 11.7 summarizes our findings and gives an outlook to identified further research challenges.

## 11.2 Context Requirements of Assistive Systems

An assistive system designed to support a person in carrying out activities in the physical environment must recognize where the user is doing what, how, and by what means. Therefore, it is necessary that the assistive system has access to a suitable abstraction of the user’s context: a personalized context model, which can be used to characterize the user’s situation and his interactions with elements of his environment.

At a meta-level [30] different but related context types can be identified that influence the setup process and the ongoing interaction between a user and the assistive system [20]. The manifestation of these context types must be *configured or learned* for every individual user installation during the *setup process of the assistive system* [see (1,2) in Fig. 11.1] and stored as his *personalized context model*. To be able to assist the user in his activities, he is to be *monitored* according to his personalized context model. The *support* opportunities depend essentially on the quality of this context knowledge [see (3, 4) in Fig. 11.1]. The following provides a brief overview

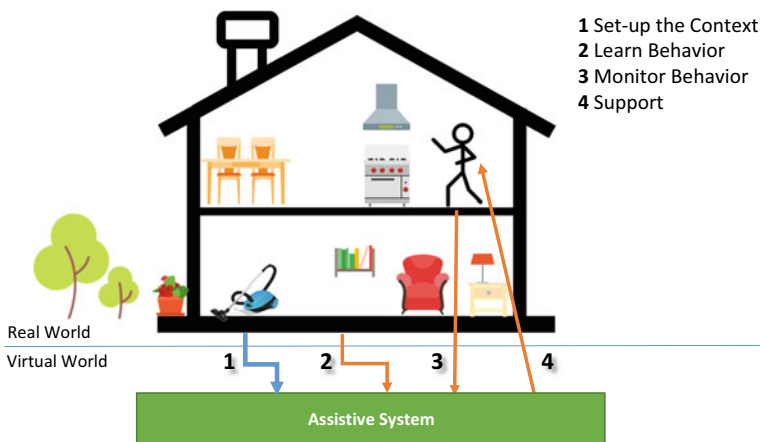


Fig. 11.1 Significance of user’s context for assistive systems

of the elements of a context model, their configuration effort, and the options for automatically monitoring user interactions with these elements.

In the real world, a person moves spatially in a certain environment, such as through rooms in an apartment or a house. The user also has special personal characteristics and abilities that can be valuable for the assistive system (e.g., blood pressure, motion profile). When setting up an assistive system, it is necessary to provide the assistive system with a model of the user's *personal* and *spatial context*, including locations, areas and their connections, and the equipment. Since the spatial environment of a user rarely changes, it can be described in the form of a floor plan including also the fixed equipment. Later, during the assistance process, it is relevant to the assistive system where the person is spatially located and how well the person is doing. Vital parameters, the position, and the movement of the person can already be monitored automatically well by HAR-Systems using built-in sensors, wearable sensors, or video observations.

It becomes more difficult for the assistive system to recognize what activity the user is performing *precisely* within his spatial environment. This ability is required both to learn the typical user behavior and provide fine-grained support if he is unable to repeat a previous behavior.

The user's interaction with resources, that make up the equipment of his spatial contexts, such as devices, applications, items, or furniture, can be used as a basis for activity recognition. Setting up the assistive system means that it is necessary to provide a model of the user's *environmental context*, including resources with their interaction possibilities and instructions. As a user can interact with a lot of different, sometimes complex resources performing his activities, the process of model construction can become very time-consuming and tedious. This is especially the case when it comes to model non-smart devices or web applications with all their user functionalities and related user instructions.

Later, during the *assistance process*, it is relevant for the assistive system to recognize what and when interactions between user and resources really take place. Only smart devices are able to provide information on their current status to the assistive system via a HAR-System. The interaction with non-smart devices still causes problems or can hardly be grasped or monitored automatically at all.

In the following, we show how HBMS-System fulfilled these context requirements. We present the context meta-model of the HBMS-System, our approach to set up personal context models in HBMS-System and derive enhancements based on our experiences regarding configuration effort and activity recognition.

### 11.3 The HBMS-System and HCM

Over the last ten years, the authors have been working on an assistive system project called *Human Behavior Monitoring and Support (HBMS)* [28]. It aims to actively support people in their daily activities (e.g., morning procedure) including also the use of their preferred web applications (e.g., e-Banking procedure). Giving support

means in HBMS to help people to remember how they once performed a particular activity by reactivating already existing memory anchors. A memory anchor is a simple stimulus that influences the state of mind. By providing retrieval cues, it is easier to remember experienced situations (cued recall) [41]. Thus, HBMS supports the autonomy of a person with decreasing memory.

In HBMS, this support functionality is based on conceptualizing a person’s episodic memory, establishing a model of that knowledge, and using that model for support. As described in Sect. 11.2, human actions take place in a physical environment. Therefore, the episodic memory of a person refers to an abstraction of his or her physical environment and user support affects again the physical environment. The personal user context is described in HBMS in the form of a suitable context model, the *Human Cognitive Model (HCM)*, building the core of the HBMS-System (see Fig. 11.2).

The user’s episodic knowledge (behavior context) as well as information about user’s environmental resources, social and personal situation, and location is used and contained in the HCM. Figure 11.3 shows a simplified meta-model of HCM. Particularly, high demands are placed on the description of the personal environmental context, including devices, household appliances, and (web) applications used [27]. The HBMS-System was realized as a model centered architecture [24]. A personalized user context model is set up in the HBMS-System in the form of models, each of which is formed tool-assisted (HCM-L Modeller) with the means of the domain-specific modeling language *HCM-L (Human Cognitive Modeling Language)* [23, 25]. It preserves knowledge in human-readable as well as in computer-readable representation form.

As shown in Fig. 11.2, when setting up the HBMS-System (1) information about the assisted person and his social context and (2) information about the user’s envi-

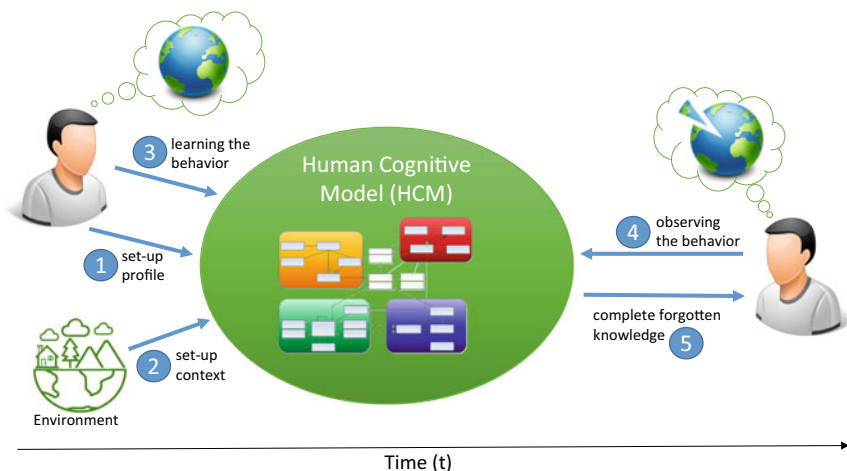


Fig. 11.2 Overall idea of the HBMS-system

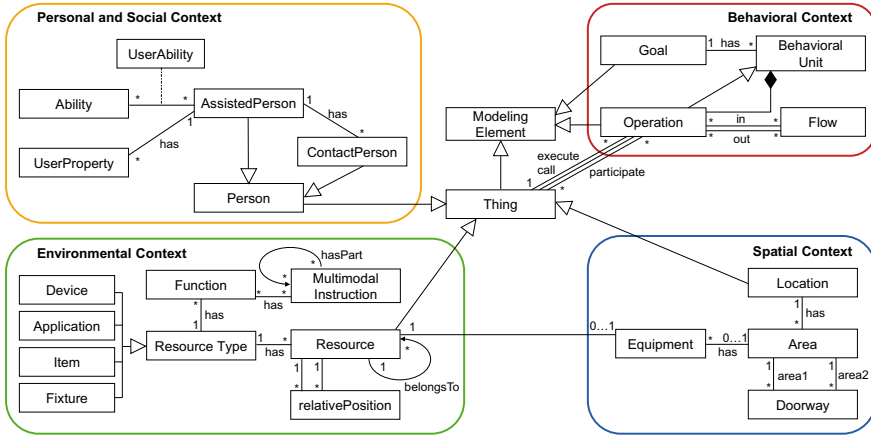


Fig. 11.3 HCM-L meta-model (excerpt from [27])

ronment and spatial context have to be added. The HBMS-System can *learn* the typical behavior (*individual episodic memory*) of the person by manual setup or by observation (3). HBMS-System is interoperable with different open HAR-Systems [36]. On demand, the HBMS-System uses this knowledge to *compensate gaps in the episodic memory* of the supported person by (4) *observing* the behavior of a person and (5) providing smart *advice and support*. Reasoning over the HCM allows the HBMS-System to predict user actions and to guide him [1].

The first evaluation of HBMS-System revealed in particular the following weaknesses:

- (1) When setting up the HBMS-System (steps 1 and 2 in Fig. 11.2), the user’s context has to be modeled manually or taken from the used HAR-System to build up the user’s personalized context model. Particularly, considerable effort is needed to model user’s resources and their functionality and handling like for household appliances and other devices [40] and web applications, as this data is not available via conventionally HAR-Systems. Thus, large parts of user manuals (instructions on how to operate the resources) have to be *captured manually during system setup* (step 2 in Fig. 11.2). This leads us to research question 1:

- *RQ1: How is it possible to improve the setup of personalized context models for assistive systems in order to keep the manual effort as low as possible?*

- (2) The HBMS-System provides interfaces for conventional HAR-Systems to be able to integrate information about users’ activities from various systems [36]. This activity information is transformed into the personalized context model (a) during the learning phase (step 3 in Fig. 11.2) to train behavior which should be supported later on and (b) during the support phase (step 4 in Fig. 11.2), where ongoing behavior is monitored step-by-step. The activity recognition results of



the used HAR-Systems concerning fine-grained user interactions with devices or web applications are not satisfying for the intended behavioral assistance [40]. For example, to pick the cutlery out of a cupboard and place it on a table would need several sensors on the cupboard, the cutlery itself and the table to be able to detect it as fine-grained as needed. It would be possible to reach better results by using depth cameras and videos, but the private home environment is too sensitive to use these technologies and end users are hardly willing to accept such solutions [2]. The need for better recognition of fine-grained user interactions with environmental resources (steps 3 and 4 in Fig. 11.2) is therefore still present in HBMS. This leads us to research question 2:

- *RQ2: How is it possible to improve the fine-grained activity recognition capability of assistive systems?*

To answer RQ1, we investigated if the semantic markup of devices and web application functions can help us to overcome the high effort during HBMS setup. Figure 11.4 shows the main idea of our approach which simplifies the construction of a personalized context model: Semantic markup data of web applications (e.g., the e-Banking application of the user) and devices (e.g., the vacuum cleaner, washing machine, mower of the user) as additional sources to set up a personalized context model (step 2a in Fig. 11.4). Nevertheless, there remain still things in the environment, such as items or fixture, which have still to be included manually (step 2b in Fig. 11.4).

To answer RQ2, we investigated if the use of *personalized and adaptive user clients can improve the activity recognition capability of assistive systems with regard to the detection of fine-grained activities*. Figure 11.5 shows the main idea of our solution: (4a) designates the sector, where behavior observation already worked well

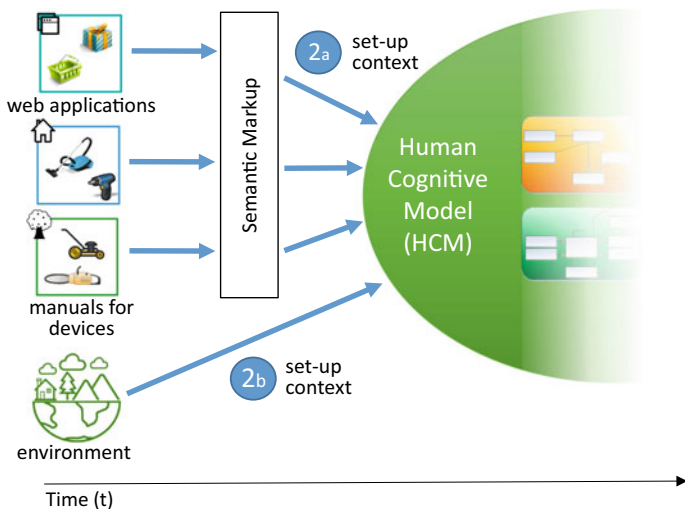


Fig. 11.4 System setup with semantic markups

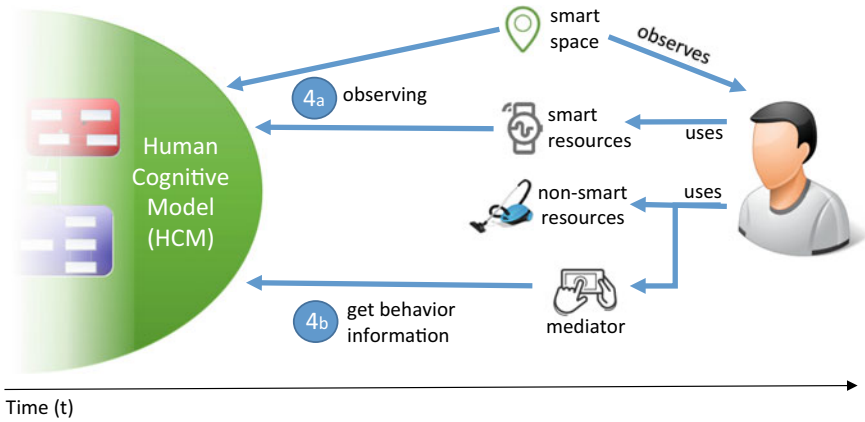


Fig. 11.5 Personalized and adaptive user clients

in HBMS. Information about the movements of a person is observed via the smart space, e.g., sensors which detect where the person is at a certain moment. Information from smart devices such as smartphones or smart watches can be easily included as well. (4b) in Fig. 11.5 shows our extension, where interactions with non-smart devices are communicated via an additional mediator, which is included in a multimodal user client of the HBMS application.

Sections 11.5 and 11.6 describe in detail how we have proceeded in answering these research questions. The next section discusses related work we need as a basis for this work.

## 11.4 Related Work and Possible Solutions

This section summarizes related work to join an assistive system with the semantic web in order to simplify the construction of the user’s context model and to improve its activity recognition capability.

### 11.4.1 Assistive Systems

The term *assistance system*, as it is understood in this chapter, is intended to provide *just-in-time activity guidance for people as they complete their activities*. A smart home is a residential home setting augmented with a diversity of multimodal sensors, actuators, and devices where this cognitive assistance can be given based on ICT services and systems. By monitoring environmental changes and inhabitants’ activities, an assistive system can process perceived sensor data, make timely decisions, and

take appropriate actions to assist an inhabitant to perform activities, thus extending their period of time living independently within their own home environment. In their survey [29], Nie et al. summarize the most recent smart home projects, and they give a classification of the main activities considered in smart home scenarios and review the different types of sensors that can be used to monitor these activities. In general, to achieve this objective, a bottom-up, sensor-centric approach is used covering the following levels:

- (1) *Monitoring*: Sensors monitor inhabitant behavior and their situated environment in real time and dynamically fuse and interpret the multiple modalities of signals. To monitor inhabitant behavior and environmental changes, visual/audial sensing facilities and networked sensor technologies can be used. Sensors can be attached to an actor under investigation (wearable sensing) for recognizing physical movements or to objects that constitute the activity environment (dense sensing) for recognizing human-object interactions [4, 29].
- (2) *Activity Recognition*: To recognize inhabitants' activities, data-driven and knowledge-driven approaches are applied [6]. Chen et al. [4] summarizes activity recognition approaches in correspondence with relevant sensors and monitored activities.
- (3) *Assistance*: Provide assistance to help the inhabitant to perform the intended activity based on recognized activities. Rafferty et al. [34] move from a sensor-centric approach for activity recognition to a top-down approach for *intention recognition*, where the inhabitants intended goals are the focus of the assistance system. The HBMS-System applies this approach too.

### 11.4.2 Semantic Markup and Web Ontologies

The semantic web offers a large variety of technologies for semantic data enrichment and semantic interoperability [3]. Jovanovic and Bagheri [18] summarize some of them in a technology stack for semantically rich e-commerce applications by dividing them into six layers:

1. **Syntactic interoperability** by using *standard data-exchange formats* (e.g., JSON, XML, RDF) and *specifications for embedding semantic markup in web pages* (e.g., Microformats, Microdata, RDFa, JSON-LD);
2. **Schema-level semantic interoperability** by using *vocabularies* (e.g., Schema.org, Open Graph Protocol, Friend of a Friend);
3. **Product identity resolution** by using *strong product identifiers* (e.g., GTIN, UPC, EAN, ISBN);
4. **Value-level semantic interoperability** by using *product catalogues/ontologies*;
5. **Advanced data search and manipulation** by using *semantic technologies* for these functionalities (e.g., RDF data storage in triple stores, SPARQL data query, and manipulation language);

6. **Improved e-commerce experience** by creating intelligent e-commerce applications (e.g., product recommendation apps).

We focus on the *second of these layers*, the *schema-level semantic interoperability*, as it comprises vocabularies for describing things on the web. These vocabularies allow for establishing syntactic and semantic interoperability. Within this layer, several ontologies for marking up things on the web exist in a large range of domains. To solve our research questions RQ1 and RQ2 formulated in Sect. 11.3, we investigated those ones, which seemed to be most promising to describe user functionalities and the handling of devices and web applications semantically.

As a source to identify candidates for our investigations, we used though not only the Linked Open Vocabularies (LOV) catalogue [22] and bundled promising candidates into the following domain categories, showing some similarities with our device and web application domain. The following ontologies have been identified as possible candidates for being a basic ontology for semantically marking up devices, user manuals, and web applications in a first step.

- **Bibliography Ontologies:** *FaBiO* [10] the “Functional Requirements for Bibliographic Records”-aligned Bibliographic Ontology is an ontology for describing publications and that contain or are referred to by bibliographic references.
- **Product and e-Commerce Ontology:** [18] discusses the advantages of using the semantic web for e-commerce. *eClassOWL is used* for describing the types and properties of products and services on the semantic web [9] and is used in combination with *goodRelations* [15], which covers commercial aspects of offers and demand, e.g., prices, payment, or delivery options. *Provoc* [33], as a product vocabulary, extends *goodRelations* for company hierarchies and production aspects.
- **Internet of Things (IoT) Ontologies:** *IoT-O* [17] is a core domain “Internet of Things” ontology, which is composed of different modules that are based on already existing ontologies [35]. *SAREF* [37] is an ontology in the smart appliances domain.
- **Web Service Ontologies:** *WSMO-Lite* Ontology [44] is a lightweight approach to semantic annotation of web service descriptions [43].
- **Geographic Information Ontologies:** There exists an OWL representation of *ISO 19115* (Geographic Information—Metadata) concerning about ordering instructions [32].
- **Web Resources/SEO:** *Schema.org* provides a single schema for structured data on the web, and its’ vocabularies include a large variety of domains (e.g., health and medicine, persons, places, products, organizations, events, actions, creative work media-objects, or recipes) [38]. *Dublin Core Schema* [8] is a small set of vocabulary terms that can be used to describe web resources like video, images, web pages, etc., and physical resources such as books, CDs, or artworks. *Schema.org* is edging *Dublin Core* out because it’s being solely and specifically created for SEO purposes by actual search engine operators [16].
- **Social media Ontologies:** The *Open Graph Protocol (OGP)* [31] allows a basic object description (e.g., Web site, image, music, video) with metadata. *OGP* is

currently used by Facebook, Google, and Mixi. *Friend of a Friend (FOAF)* [11] concentrates on relations between agents (persons, groups, organizations) and documents.

- **Operating instructions:** no specific vocabulary for operating instructions could be identified in *LOV-catalogue*; [42] presents a markup of medical guidelines which is not applicable to our requirements; and the *Learning Resource Metadata Initiative (LRMI) ontology* [21] is a collection of classes and properties for markup and description of educational resources. The specification builds on the extensive vocabulary provided by Schema.org and other standards.

In [12, 39], we specified the *concepts needed to describe devices and web applications semantically* and *evaluated* the most promising ontology candidates. *Schema.org* seems to be the best approach, due to the concepts overlap, its wide distribution, its extensibility, and the support of search engine optimization. In Sect. 11.5, we show in more detail how to semantically describe devices and web applications, their components and functions, and their handling using the schema.org vocabulary.

## 11.5 Semantic Markup of Devices and Web Applications

Section 11.3 describes the problems that have occurred with our assistive system HBMS when setting up the user environment and observing interactions with non-smart devices and web applications. We hypothesized that *semantic markup* could help to overcome these problems.

Information regarding the general handling of a device or a web application represents domain knowledge. It is independent of a certain user and usually described by the manufacturer or developer in the form of a user manual.

For devices, manufacturers provide manuals to inform a user how to prepare, use, transport, store, and maintain them. These manuals are typically described in the form of multipage, multilingual, detailed documents, made available for the user by hardcopy or online. Sometimes explanatory videos are additionally offered on social media channels. These manuals are used in the following to describe devices semantically.

In contrast to devices, web applications usually offer context-sensitive user manuals online. Interactive user elements can be automatically identified in HTML code, but the semantics of what actions are triggered when interacting with these elements are not included in HTML code. Schema.org metadata is already used on web pages to mark products, events, organizations, or even some actionable elements, but it is not used to semantically describe the *entire user interface of a web application*.

This section shows how to use schema.org to semantically describe properties, functions, and operating instructions of devices and web applications. A web-based tool called “Schemator” has been developed yet [12] to semi-automatically create appropriate schema.org semantic markups for web applications. The semantic description created by the manufacturer or developer in such a manner can be down-

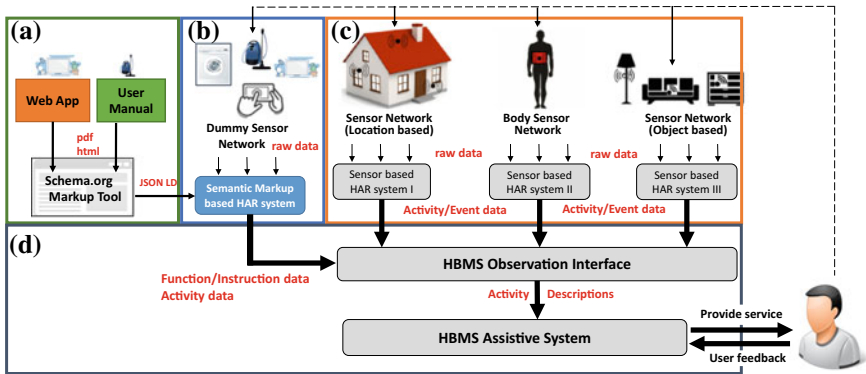


Fig. 11.6 Semantic markup interoperability scenario

loaded from the web in JSON-LD format (see Fig. 11.6a). This enables interoperability with various assistive systems, and the environmental context knowledge needed for a specific user can be imported from the web into the personalized context model of the assistive system.

To improve the activity recognition power and interaction granularity, we propose personalized and adaptive user clients simulating sensor data (see Fig. 11.6b). This approach will be detailed in Sect. 11.6. Via the HBMS Observation Interface, this “semantic based” HAR can be integrated like location or object-based sensor-based HARs [4] (see Fig. 11.6c) to work interoperable with the assistive system and improve its activity recognition capability (see Fig. 11.6d).

### 11.5.1 Semantic Markup of Web Applications

To semantically describe the interaction possibilities of a user with a web application (see Fig. 11.6a), we examined several web applications and modeled web applications at the meta-level using suitable *schema.org* types. Figure 11.7 sketches the result of this process.

A web application can be described using *WebApplication*. Each page of the web application can be described using *WebPage* and its interactive elements using *WebPageElement*. As all these types are subtypes of *CreativeWork*, properties are available to describe *identifier*, *name*, and *description*. *MediaObjects* and more specific types like *AudioObject*, *VideoObject*, and *ImageObject* can be linked too using the property *associatedMedia*. What triggers the interaction with a *WebPageElement* can be described using the property *hasPotentialAction* referencing an *Action*. Actions were introduced in *schema.org* to describe the ability to perform a certain operation on a *Thing*. Each Action has a *name*, a *description*, and a *result*. Latter is of type

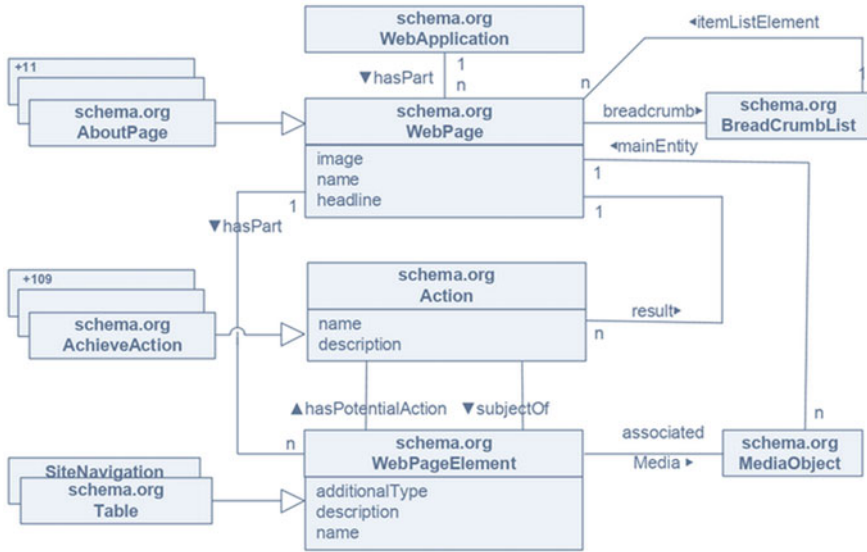


Fig. 11.7 Meta-model for web application markup

*CreativeObject* (and thus, e.g., again a *WebPage*). Several specific *Action* types like *AchieveAction*, *SearchAction*, *CreateAction*, and more are predefined in schema.org.

The schema.org markup tool “Schemator” (see Fig. 11.6a) supports generating schema.org metadata of web applications semi-automatically: In a first step, it reads HTML and JavaScript elements of a given web page. In the second step, it provides a user-friendly interface to add semantic information that cannot be generated automatically. Compared to conventional schema.org tools, only the data necessary in this context is requested from the tagging user. The schema.org data of the marked-up web application is centrally stored and can be downloaded afterward in JSON-LD format. Figure 11.8 shows an excerpt of the JSON-LD code describing an example web application.

### 11.5.2 Semantic Markup Data of Devices

While human readers understand the handling of a good user manual mostly at a glance, an assistive system needs extra information. In this section, we show how to describe user manuals semantically including functions, problem situations, warnings, and instructions.

Although users are most widely interested in device’s core functionalities to support them in their activities, also “support” functions (installation, maintenance) have to be carried out and described in manuals. After having examined several manuals, we figured out the needed information for manual’s semantic comprehension and

```

"@context"           : "http://schema.org",
"@id"               : "https://developer.mozilla.org/en-US/docs/Web/HTML/Element/html?1",
"type"              : "WebApplication",
"url"               : "https://www.basicwebshopexample.com",
"name"              : "Basic Webshop Example",
"creator"           : "John Doe",
"applicationCategory" : "Webshop",
"about"             : "This is an basic annotation example",
"hasPart":
{
  "@type"           : "WebPage",
  "@id"             : "https://developer.mozilla.org/en-US/docs/Web/HTML/Element/body?1",
  "hasPart" :
  {
    "@id"           : "https://developer.mozilla.org/en-US/docs/Web/HTML/Element/main?1",
    "@type"         : "WebPageElement",
    "hasPart"       :
    {
      "@id"           : "https://developer.mozilla.org/en-US/docs/Web/HTML/Element/i?1",
      "@type"         : "SiteNavigationElement",
      "name"          : "i.fa.fa-shopping-bag",
      "additionalType": "Brand",
      "description": "The brand logo of this webshop. Click on it to return to the start site.",
      "potentialAction" :
      {
        "@type"       : "Action",
        "name"         : "onclick.navigateToHome",
        "description" : "Return to start site.",
        "result"       :
        {
          "@type" : "WebPage",
          "@id"   : "https://developer.mozilla.org/en-US/docs/Web/HTML/Element/body?2"
        }
      }
    }
  }
}

```

**Fig. 11.8** Excerpt of the JSON-LD code describing a web application

modeled devices, their functionality and associated instructions at the meta-level. As in Sect. 11.5.1 for web applications, we reused schema.org as far as possible for the semantic markup of devices.

Figure 11.9 sketches the result of this process and focuses on those schema.org classes and properties, which are suitable for this purpose. The similarity with the meta-model for the description of web applications in Fig. 11.7 is obvious.

In contrast to the latter, we also used the extension mechanism of schema.org. In the following, we focus on the identified *existing* (shadowed in Fig. 11.9) and *extended* schema.org classes and properties (white in Fig. 11.9) to semantically markup devices and their handling. As the e-commerce schema from the goodRelations project [15] has been integrated into schema.org in 2012, it is easy to express structured data about products and related facts with schema.org vocabulary. Thus, a device can be easily described as a *Product*, with the properties *name*, *description*, *identifier*, *category*, and *image*. For the specialization of a product into different models, schema.org offers the subclass *ProductModel* and the property *successorOf* to tag developments of a product over time. Complex devices can be mapped to *Product* as well, as *Product* includes a property *isAccessoryOrSparePartFor* referring to another product (or multiple products) for which this product is an accessory or spare part.

Every *Thing* and thus, also *Product* can have a set of *potentialAction* of type *Action*. This property references an idealized action in which this thing would play an “object” role. There exist many action subclasses, and *useAction* seems to be well suited for tagging **Core Functions of a device**. A fitting *Action* subclass to map



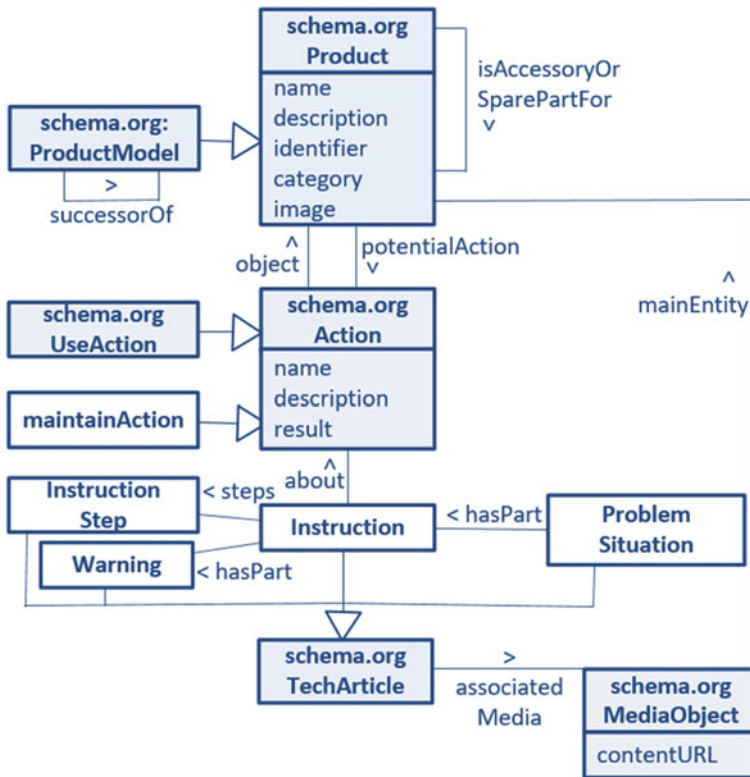


Fig. 11.9 Schema.org excerpt (colored fields) and our extensions (white)

**Support Functions** has not been defined yet. Thus, a specialization of *Action* to *maintainAction* is helpful to tag them.

Manuals can be considered as creative work. Schema.org offers *CreativeWork* with a lot of subclasses to tag web content like *Article*, *Book*, *MediaObject*, *Movie*, *Recipe*, *Web site*, and more. Although no schema.org subclasses for manuals have been defined yet, *TechArticle*, a subclass of *Article*, can be used to map the concepts of *Instruction*, *Instruction Step*, *Warning*, and *Problem Situation*, which can be found in almost every user manual. Thus, a specialization of *TechArticle* into *Instruction*, *InstructionStep*, *Warning*, and *ProblemSituation* and the extension of *Instruction* with the property *steps* do make sense using *additionalType*.<sup>1</sup> The existing property *hasPart* allows to indicate other *CreativeWorks* that are parts of this *CreativeWork*, so it allows to map the interrelationship between *ProblemSituation*, *Instruction*, and *Warning*. *MediaObject* and more specific types like *AudioObject*, *VideoObject*, and *ImageObject* can be associated with *TechArticle* using the property *associatedMedia*.

<sup>1</sup> See, e.g., <http://sdo-schemaorgae.appspot.com/TechArticle>.

As every *CreativeWork* can have a subject matter of the content using the property *mainEntity*, this property can be used to assign it to a *Product* or *Action* tagged before.

Like sketched in Fig. 11.8 for a web application also the semantic description of a device can be represented in JSON-LD format. During HBMS-System setup, this device knowledge can be imported from the web into the personalized context model.

### 11.6 Personalized and Adaptive User Clients

This section uses an example to explain how semantic data is integrated into the HBMS-System and how this information is used for active support. As we see in Fig. 11.10, the assumption for the following example is, that manufacturers semantically marked up their devices (e.g., all types of vacuum cleaners) online using the “Schemator” tool (see Sect. 11.5).

During HBMS-System setup and customization, the personalized user context has to be defined (see Sect. 11.3). This covers, e.g., also the description of the functionality and handling of a vacuum cleaner of the supported user. As mentioned, the manual resource definition and update of the environmental context is a considerable effort we wanted to reduce. Thus, the automatic import of the semantic data about the given type of vacuum cleaner into the personalized context model highly facilitates this process.

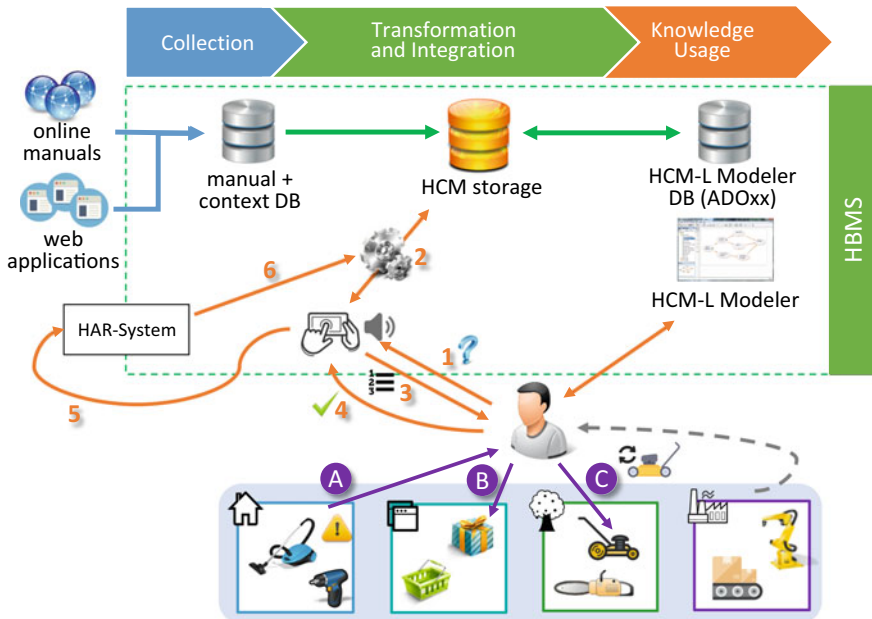


Fig. 11.10 Personalized and adaptive user clients in HBMS

The data is taken from the web, transformed, and integrated into the HBMS-System data stores mapping tagged elements to HCM-L *resource types*, HCM-L *functions* or *multimodal instructions* (see Fig. 11.3). The HCM-L Modeler [1] enables to visualize the imported information.

During the HBMS-support process, the user is guided on demand through behavioral units (BUs) and their operations. These BUs can be indoor or outdoor, at home or in business. We assume in our example the user wants to clean up his apartment and uses his vacuum cleaner.

If one operation of a BU includes the use of a special resource functionality or gives a warning [e.g., (A) in Fig. 11.10, dust bag of the vacuum cleaner has to be replaced], (1) the *user contacts* the HBMS-System, which (2) *retrieves* the demanded *knowledge* regarding this vacuum cleaner and (3) *passes* the *appropriate information* via a client (e.g., a tablet or a voice assistant) to instruct the user step-by-step. After each instruction step, (4) the *user feedbacks* the HBMS-System about the current execution status (e.g., via click or voice assistant). The HBMS-System (5) *simulates sensor data* based on this user feedback and sends it to the related HAR-System. From the point of view of the HAR-System, it seems as if the data had come from a genuine sensor. Thus, (6) *activity data* is handled by the HBMS-System as well as if it were recognized via a real sensor. It is *included* in the HBMS knowledge base and used in succeeding operations. (2)–(6) are now repeated *until the problem is solved* and the warning is gone.

The steps (1)–(6) can also be applied for web applications or web pages [see (B) in Fig. 11.10] which are *semantically marked up*, as the information about the steps needed to use the web application or web page is included in the HCM-storage such as other data.

Moreover, automated mapping from old to renewed devices is possible. If a device is replaced by a new one or a new device is added [e.g., (C) in Fig. 11.10], the update of the manual goes the same “collection, transformation, integration” way until the manual information is contained in the knowledge base and related to BUs using the former device. Consequently, it can be used for support.

## 11.7 Summary and Outlook

This contribution combines assistive systems with the semantic web using HBMS-System as an example, in order to simplify the creation of a personalized context model during system setup and to improve the system’s activity recognition capabilities.

This approach makes interaction possibilities and functions of devices and web applications semantically *understandable* for the HBMS-System and *interoperable* with its environmental context model. We show how to semantically describe devices and web applications including their functionalities, results, and user instructions and propose to use the vocabulary of schema.org. Since schema.org is very extensive, we define a vocabulary subset, but also a small extension of schema.org to fill in

gaps. The result is presented in the form of two meta-models. According to these meta-models, devices and web applications can be marked up and represented in JSON-LD format on the web. In this way, “semantic manuals” are created. A tool therefor (‘Schemator’) is in development.

Furthermore, this contribution explains how *personalized and adaptive* HBMS *user clients* and the power of the *HBMS environmental context* model can be used to *bridge* an existing *activity recognition gap*. User feedback after each instruction step is included in the HBMS-System by simulation of sensor data according to this user feedback. We are currently working on the implementation of the capture, transformation, and integration process in the HBMS-System.

Although we have used a domestic example, the idea of semantic manuals is transferable to other areas such as software applications or machines in production halls and their manuals. Describing industrial information models with ontologies and constraints is an occurring research topic [19]. To provide active assistance for such processes in the industrial context (e.g., for manufacturing processes), semantic markup of *industrial manuals* is a promising approach. In addition, *software applications* could benefit from semantic manuals. Consequently, further development of the HBMS-System will *focus on these domains*.

Even better tools are needed to facilitate markup for manufacturers. A natural language approach that generates the required structured metadata from text documents on the web will be helpful for manufacturers and is another interesting area of research.

Furthermore, the *Internet of Things (IoT)* world offers a large variety of open issues for the research community. First adaptations of the context model toward IoT manufacturing processes and privacy concerns are already in progress [26]. This also includes approaches to supporting interoperability through ontologies [7]. Working with heterogeneous intelligent systems in assistive systems remains a challenge where the use of ontologies to standardize and support semantic interoperability would bring great benefits to ongoing projects [5]. Furthermore, the ontological foundation of the HBMS meta-model in a basic ontology [14] by the research team should improve the quality of conceptual modeling languages and models.

## References

1. Al Machot F, Mayr HC, Michael J (2014) Behavior modeling and reasoning for ambient support: HCM-L modeler. In: Hutchison D et al (eds) Modern advances in applied intelligence. International conference on industrial engineering and other applications of applied intelligent systems, IEA/AIE 2014, Kaohsiung, Taiwan, proceedings, Part II. Springer International Publishing, Cham, pp 388–397
2. Arning K, Zieffle M (2015) “Get that camera out of my house!” Conjoint measurement of preferences for video-based healthcare monitoring systems in private and public places. In: Geissbühler A et al (eds) Inclusive smart cities and e-health. 13th international conference on smart homes and health telematics, ICOST 2015, Geneva, Switzerland, LNCS 9102. Springer, Cham, pp 152–164

3. Chan LM, Zeng ML (2006) Metadata interoperability and standardization—a study of methodology part I: achieving interoperability at the schema level. *D-LIB Mag* 12(6)
4. Chen L, Hoey J, Nugent CD, Cook DJ, Yu Z (2012) Sensor-based activity recognition. *IEEE Trans Syst, Man, and Cybern-Part C* 42(6):790–808
5. Chen L, Nugent CD, Okeyo G (2014) An ontology-based hybrid approach to activity modeling for smart homes. *IEEE Trans Hum-Mach Syst (THMS)* 44(1):92–105
6. Chen L, Nugent CD, Wang H (2012) A knowledge-driven approach to activity recognition in smart homes. *IEEE Trans Knowl Data Eng* 24(6):961–974
7. Daniele L, Solanki M, den Hartog F, Roes J (2016) Interoperability for smart appliances in the IoT world. In: Groth P et al (eds) *The semantic web—ISWC 2016*, LNCS 9982, Springer, pp 21–29
8. Dublin Core Metadata Initiative (DCMI) (2019) URL: [goo.gl/gy4P9Q](http://goo.gl/gy4P9Q). Last access: 11.3.2019
9. eClassOWL—The Web Ontology for Products and Services (2007) URL: [goo.gl/OgEJFg](http://goo.gl/OgEJFg). Last access: 11.3.2019
10. FaBio (2016) The FRBR-aligned Bibliographic Ontology. URL: [goo.gl/CW9QfL](http://goo.gl/CW9QfL). Last access: 9.5.2017
11. Friend of a Friend (FOAF) (2015) URL: [goo.gl/PH5YBR](http://goo.gl/PH5YBR). Last access: 11.3.2019
12. Friesser J (2019) *Semantische Aufbereitung von Benutzeroberflächen von Webapplikationen für die aktive Assistenz mit HBMS*, diploma thesis
13. Geisberger E, Broy M (eds) (2012) *agendaCPS: Integrierte Forschungsagenda Cyber-physical systems*, vol 1. Springer
14. Guizzardi G, Falbo R, Guizzardi RS (2008) Grounding software domain ontologies in the unified foundational ontology (UFO): the case of the ODE software process ontology. In: M Lencastre et al (eds) *11th Ib. WS on RE and SW Environments*, pp 127–140
15. Hepp M (2008) *GoodRelations: an ontology for describing products and services offers on the web*. In: Gangemi A, Euzenat J (eds) *Knowledge engineering: practice and patterns*, vol 5268. Springer, Berlin, Heidelberg, pp 329–346
16. Hisham A (2015) Dublin core versus Schema.org: a head-to-head metadata comparison. URL: [goo.gl/btkd91](http://goo.gl/btkd91). Last access: 11.3.2019
17. IoT-O (2017) a core domain Internet of things ontology. URL: [goo.gl/CH1cdX](http://goo.gl/CH1cdX). Last access: 11.3.2019
18. Jovanovic J, Bagheri E (2016) Electronic commerce meets the semantic web. *IT Prof* 18(4):56–65
19. Kharlamov E et al (2016) Capturing industrial information models with ontologies and constraints. In: Groth P et al (eds) *The semantic web—ISWC 2016*, LNCS 9982, Springer, pp 325–343
20. Kofod-Petersen A, Cassens J (2006) Using activity theory to model context awareness. In: Roth-Berghofer TR, Schulz S, Leake DB (eds) *Modeling and retrieval of context*. MRC Workshop 2005, Edinburgh, UK, revised selected papers, LNCS 3946. Springer, Berlin, pp. 1–17
21. Learning Resource Metadata Initiative (LRMI) (2018) URL: [goo.gl/APM6IE](http://goo.gl/APM6IE). Last access: 8.3.2019
22. Linked Open Vocabularies (LOV) (2019) URL: [goo.gl/JWhNDU](http://goo.gl/JWhNDU). Last access: 11.3.2019
23. Mayr HC, Al Machot F, Michael J, Morak G, Ranasinghe S, Shekhovtsov V, Steinberger C (2016) HCM-L: domain-specific modeling for active and assisted living. In: Karagiannis D, Mayr HC, Mylopoulos JP (eds) *Domain-specific conceptual modeling. Concepts, methods and tools*, vol 5. Springer, Cham, pp 527–552
24. Mayr HC, Michael J, Ranasinghe S, Shekhovtsov VA, Steinberger C (2017) Model centered architecture. In: Cabot J et al (eds) *Conceptual modeling perspectives*. Springer International Publishing, Cham, pp 85–104
25. Michael J, Mayr HC (2013) Conceptual modeling for ambient assistance. In: Ng W, Storey VC, Trujillo J (eds) *Conceptual modeling—ER 2013*. 32th International conference on conceptual modeling. Hong-Kong, China, LNCS 8217. Springer, pp 403–413
26. Michael J, Koschmider A, Mannhardt F, Baracaldo N, Rumpe B (2019) User-centered and privacy-driven process mining system design. To appear in *CAiSE Forum*

27. Michael J, Steinberger C (2017) Context modelling for active assistance. In: Proceedings of the ER Forum 2017 and the ER 2017 demo track co-located with ER 2017, CEUR workshop proceedings (CEUR-WS.org), pp 221–234
28. Michael J, Steinberger C, Shekhovtsov VA, Al Machot F, Ranasinghe S, Morak G (2018) The HBMS story. Enterprise modelling and information systems architectures. *Int J Concept Model* 13:345–370
29. Ni Q, Garcia Hernando AB, Pau de la Cruz I (2015) The elderly's independent living in smart homes: a characterization of activities and sensing infrastructure survey to facilitate services development. *Sensors* 15(5):11312–11362
30. Object Management Group OMG (2016) Meta Object Facility (MOF) Core, URL: <http://bit.ly/OMG-MOF>, Version 2.5.1. Last accessed 11.3.2019
31. Open Graph Protocol (OGP) (2014) URL: [goo.gl/51KfLM](http://goo.gl/51KfLM). Last access: 11.3.2019
32. OWL Representation of ISO 19115 (Geographic Information-Metadata) (2014) URL: [goo.gl/ygbQZ7](http://goo.gl/ygbQZ7). Last access: 11.3.2019
33. Provoc—Product Vocabulary (2016) URL: [goo.gl/HGm3u1](http://goo.gl/HGm3u1). Last access: 11.3.2019
34. Rafferty J, Nugent CD, Liu J, Chen L (2017) From activity recognition to intention recognition for assisted living within smart homes. *IEEE Tr. Hum-Mach Syst* 47(3):368–379
35. Seydoux N, Drira K, Hernandez N, Monteil T (2016) IoT-O, a core-domain IoT ontology to represent connected devices networks. In: Blomqvist E et al (eds) European knowledge acquisition workshop (EKAW 2016): knowledge engineering and knowledge management, Springer, pp 561–576
36. Shekhovtsov VA, Ranasinghe S, Mayr HC, Michael J (2018) Domain specific models as system links. In: Woo C, Lu J, Li Z, Ling TW, Li G, Lee ML (eds) Advances in conceptual modeling. ER 2018, Xian, China. Springer International Publishing, Cham, pp 330–340
37. Smart Appliances REferences (SAREF) Ontology (2013) URL: [goo.gl/Y93h64](http://goo.gl/Y93h64). Last access: 11.3.2019
38. Stavrakantonakis I, Fensel A, Fensel D (2014) Matching web entities with potential actions. In: SEMANTiCS 2014—posters & demos track. CEUR
39. Steinberger C, Michael J (2017) Semantic mark-up of operating instructions. Tech. Report. URL: <http://bit.ly/semanticMarkup>. Last access: 11.3.2019
40. Steinberger C, Michael J (2018) Towards cognitive assisted living 3.0. In: IEEE international conference on pervasive computing and communications workshops (PerCom workshops), Athens. IEEE, Piscataway, NJ, pp. 687–692
41. Strube G (ed) (1996) Wörterbuch der Kognitionswissenschaft. Klett-Cotta, Stuttgart
42. Svátek V, Růžička M (2003) Step-by-step mark-up of medical guideline documents. *Int J Med Inf* 70(3):329–335
43. Vitvar T, Kopecký J, Viskova J, Fensel D (2008) WSMO-lite annotations for web services. In: Bechhofer S (ed) The semantic web: research and applications. 5th European semantic web conference, ESWC 2008, Tenerife, Spain. Springer, Berlin, pp 674–689
44. WSMO-Lite Ontology (2013) URL: [goo.gl/5y5uxo](http://goo.gl/5y5uxo). Last access: 11.3.2019

**Part IV**  
**Open Smart Home and Service**  
**Infrastructures**

# Chapter 12

## Towards Multi-resident Activity Monitoring with Smarter Safer Home Platform



Son N. Tran and Qing Zhang

**Abstract** This chapter demonstrates a system that can turn a normal house to a smart house for daily activity monitoring with the use of ambient sensors. We first introduce our smarter safer home platform and its applications in supporting independent livings of seniors in their own home. Then we show a proof of concept which includes a novel activity annotation method through voice recording and deep learning techniques for automatic activity recognition. Our multi-resident activity recognition system (MRAR) is designed to support multiple occupants in a house with minimum impact on their living styles. We evaluate the system in a real house lived by a family of three. The experimental results show that it is promising to develop a smart home system for multiple residents which is portable and easy to deploy.

**Keywords** Multi-resident activity · Smart homes · Ambient intelligence · Voice-based annotation

### 12.1 Introduction

Smart home technology has changed the way we live significantly, making it more comfortable and joyful. The idea of ‘smart’ in smart homes originally focuses on employing ubiquitous computing to manage home appliances in order to transform an ordinary house to an autonomous system [1, 2]. Early systems have achieved notable success in commercialisation which has driven the increasing demand for practical applications, especially for activity monitoring in health care. This motivated intensive research for smarter, safer homes which not only passively sense and respond but also be able to actively and securely reason about the behaviour of

---

S. N. Tran  
ICT Discipline, The University of Tasmania, Launceston, Australia  
e-mail: [sn.tran@utas.edu.au](mailto:sn.tran@utas.edu.au)

Q. Zhang (✉)  
The Australian E-Health Research Centre, CSIRO, Level 5, UQ Health Science Building, RBWH,  
Herston, QLD 4029, Australia  
e-mail: [qing.zhang@csiro.au](mailto:qing.zhang@csiro.au)



residents through the activation of sensors. The ‘smart’ idea is therefore shifting its focus from ‘home control and automation’ [1] to ‘monitoring and assistance’ [3].

In health care domain, especially aged care, activity monitoring of residents in smart homes plays an important role as it can provide information about reflections of the residents’ well-being. Such information is critically important for the development of an intelligent system to assist the residents in the cases of urgent incidents such as falls and strokes and also to track and predict any risks of health problems in the long term. The key challenge here, however, is to select a suitable technology to obtain the needed information. Early approaches employ visual computing as rich information from visual images can be used to recognise activities with high accuracy [4–8]. Despite the effectiveness, the use of high-quality cameras raises a serious concern about privacy. In order to solve this issue, depth cameras have been used [9–11]; however, as a trade-off, depth cameras have low-quality images and are more sensitive to the occlusion problems. Moreover, the appearance of cameras makes residents uncomfortable with the feeling of being watched in their private houses. Other approaches rely on wearable sensors where information is collected from each resident for reasoning about their behaviour [12–14]. This may be the ad hoc solution for a short period; however, it cannot meet the increasing desire for smarter homes in which residents do not need to change living patterns to accommodate monitoring requirements from wearable sensors. Moreover, wearable sensors require some attention from their users, i.e. power charging and regular quality check, which is not very convenient especially for senior citizens or people with difficulties. Recently, there is increasing attention to ambient intelligence which is expected to overcome the privacy issue (as mentioned in visual-based approaches) and the obtrusive issue (as mentioned in wearable approaches). We believe that the most practical solution should entail ambient sensors. The key advantages of ambient sensors include power saving and low cost, and especially, they do not require human’s attention since they are attached on the surrounding objects. The greatest weakness of ambient sensors is they are very sensitive to noise that causes considerable difficulty for processing data and for learning a recognition model. This problem, fortunately, can be ameliorated by statistical modelling methods as what have been shown in previous works [15–21].

Another challenge for development of modern smart homes is the scalability in activity modelling, i.e. how to deal with multiple occupants. Most of previous work on activity recognition in ambient environment focuses on single resident, aiming to support independent living [22–26]. However, it is very common that homes are occupied by more than one resident, and therefore, we need smarter solutions to be able to reason about activities of multiple residents. For activity recognition in smart homes, there are two groups of the approaches: knowledge-driven and data-driven; however in the case of multi-resident, most approaches are data-driven, see [27] for a survey. In data-driven activity modelling, input data can be obtained from sensors but output data relies heavily on annotation process which is time-consuming and requires much of manual effort. A common idea for collecting activity labels is to use diaries or logging applications that requires residents to write/type in their activities [15, 28–30]. This, however, places more burden on to the shoulders of residents where

many of them, especially senior citizen, found it difficult to remember the activities they have performed in details. Another idea is to employ experimenters to guess the activities of residents in a house with the help of visualisation tool [31, 32]. The disadvantage of this method is that it requires too much manual effort while a smart system should minimise the human involvement. Besides annotation, there is also an intensive study in modelling activities of different residents. A straightforward approach is to train an activity model for each resident [17] from sensor data. In order to do this, the knowledge of ‘what sensor is activated by whom’, as known as *data association*, should be given. Such prior knowledge is difficult to obtain and even requires more labelling effort. Several approaches proposed an idea of predicting data association at the same time with predicting activity [33–35]. It has been shown that data association has a strong correlation with residents’ activities such that accurate tracking of all residents can improve the recognition performance [33]. However, it can also be reasoned that bad prediction of data association would be problematic to performance of activity recognition for multiple residents, as we can see in [33]. Alternatively, we can ‘ignore’ the data association problem by modelling the behavioural dependencies to predict the activities of multiple residents at the same time. Static models such as k-nearest neighbours, decision tree and neural networks can be used [36], but dynamic Bayesian models such as hidden Markov models (HMMs) and conditional random field (CRF) are more popular as they can present the temporal nature of the sensor data [17, 33]. Recently, deep learning techniques have been emerging as effective tools for data-driven modelling but have not been used widely for activity recognition in ambient intelligence environment.

In this chapter, we introduce a smarter safer home platform built upon ambient technologies combined with natural language processing-based techniques for annotation and deep learning for multi-resident activity modelling.

Our system is illustrated in Fig. 12.1. It addresses the first challenge of privacy and obtrusive concerns by using low-cost, energy-efficient ambient sensors for collection of behavioural information of residents. The system also addresses the second challenge of multi-resident activity modelling by (1) proposing an annotation approach based on natural language processing (NLP) and (2) employing multi-label recurrent gated units (GRU) [37], a state-of-the-art deep learning technique for multi-resident activity prediction. Especially, we design the system as portable as possible such

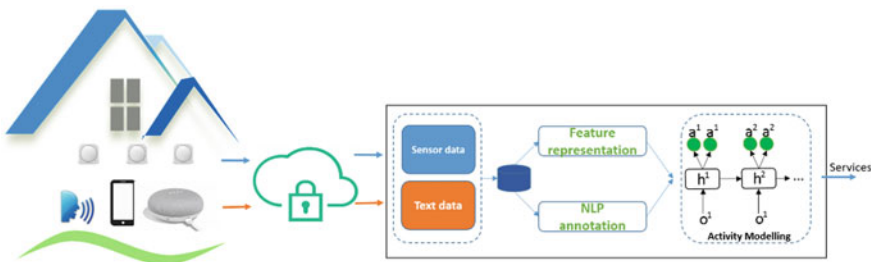


Fig. 12.1 Multi-resident activity recognition system (MRAR) in smart homes

that it can easily convert a normal home into a smart home. In practice, we deploy the system to a real house occupied by a typical family, a couple with children. We install the sensors without rearranging anything in the house. The data is collected by a gateway to send to our processing module. The activities are logged verbally by a virtual assistant where the speech is converted to text and sent to our database. Here, we process the speech-to-text data to label the activities and use the sensor data to train a recognition model.

The rest of this chapter is organised as follows. The next section discusses the related work on smart homes and on activity modelling for multiple residents. In Sect. 12.3, we present our SSH platform including the hardware and communication protocols. Section 12.4 describes the idea of using LNP and deep learning techniques for activity recognition in the SSH. In Sect. 12.5, we deploy and evaluate our SSH system in a real house. Section 12.6 concludes the chapter.

## 12.2 Related Work

### 12.2.1 *Smart Homes*

As longevity of older people increases with advances in medical therapeutics and devices, it is very likely the elderly population will prefer and want to live longer in their homes. Aside from their preferences, the cost and lack of residential care placements are additional factors that will influence families to look adopt assistive technologies to support their parents at home. As health and lifestyle monitoring technologies are becoming embedded in our daily life and Internet connectivity is becoming pervasive, the baby boomers who are readily adopting these technologies would want their homes to be smart to prolong a healthy lifestyle.

The original ‘Smart Home’ concept was proposed in the 1980s and found its application in health and ageing to support independent living of elderly people. Along with the emergence of new technology in mobile computing, smart sensors and the Internet of Things, smart homes are becoming a hot topic and are poised for strong growth in home automation, assistance and health and well-being.

The recent emergence and advances in wireless sensor, and mobile computing technologies have made lifestyle oriented monitoring easily integrated in a smart home environment. For smart home integration to be effective in supporting an elderly person at home, it needs to be able to determine if functional ability and independence, together with health, are maintained. A few smart home-like products, such as Just Checking [38], My Lively [39], and Tunstall [40], are emerging in the market place that are employing motion and movement sensors to detect daily activity and the possibility of detecting falls. There are already telehealth products such as Telemedcare [41] and Intel’s Care Innovations [42] that support health monitoring. These products, however, are more specialised for those with severe chronic illness

and are costly for general health monitoring of elderly who remain functionally independent.

However, the wide adoption and deployment of smart homes in the senior community are still elusive. We believe there are two main reasons which have limited smart home initiatives. First, the cumbersome use of technology not particularly tailored to ease of use by seniors together with the perception that technology will compromise their privacy and security. Second, there are no personalised and objective indicators of functional independence that determine the health and well-being of senior residents within their own home environment.

Motivated by these limitations, the CSIRO SSH platform was designed as a passive activity monitoring system, without the need of intervention by residents, to capture their requirements for support and services. The platform was designed to be interoperable with commercially available sensors and devices. Furthermore, the design included privacy and security considerations, ensuring informed consent of all monitoring and data collection processes. A study with prospective residents further ensured the design was user-centric and met the stated needs of residents, including that the system be flexible, low maintenance and unobtrusive [43].

### ***12.2.2 Multi-resident Activity Modelling***

In smart home, activity recognition modelling is a well-studied research area including knowledge-driven and data-driven approaches. For the knowledge-driven approach, labelled data is not necessary, yet domain and a prior knowledge of home activities must be available. For better understanding, readers should refer to an important work on this research direction of activity recognition in smart homes [44]. There is also a hybrid approach for activity modelling that combines data-driven and knowledge-driven approaches [45].

In this chapter, we focus on the emerging trend of learning/mining from large amount of data. A common problem of such data-driven approach is that acquiring heavily labelled training data in real smart home applications is usually prohibitive. This motivates our proposed voice agent-based auto-data-label approach in this chapter. For data annotation, there are two issues that challenge the development of an effective smart home, namely *intrusiveness* and *labelling cost*. The former concerns whether residents are comfortable or not while the latter concerns how much effort should be used for annotation. Early work employs cameras which can capture rich information about daily activity [4–8]. However, such approach raises strong arguments over the privacy issue, and it also requires expensive manual labelling cost as recording activities in smart homes would take a long period of time. Second approach, which is among the most popular ones, uses computer-based logging program [15]. This method requires residents to be familiar with using modern computers, and more importantly, the logging activity may interfere with the actual routines of residents when living in a smart home. Moreover, it is difficult to use such method for aged care as senior citizens may not be able to remember every

activity and their times for a whole day. In order to free residents from labelling task in [31, 32], visualisation tools are proposed. However, this still assumes a lot of human involvement where experts are hired to label activities using the visualisers. Besides, due to the noise and ambiguity in data for visualisers and the fact that all labels are generated by human's guessing it is highly likely that the labels are not perfect. Another approach has been developed recently is generating labelled data from a simulator [46]. However, it is not clear how this can help activity modelling in real smart homes.

Application of data mining techniques for multi-resident activity monitoring has been an intensive focus recently due to the increasing demand for practical smarter homes. A notable work can be referred to the CASAS project [47]. This project has been being a spotlight in the domain since 2007 with hundreds of publications along with public data and supporting tools. Similar to our work, CASAS smart homes mainly use ambient sensors, especially motion sensor for movement detection. However, different from NLP-based annotation in our system, their data annotation relies on residents' activity diary and/or experts' labelling with the help of visualisation software [31, 48]. Another similar project ARAS [15] included only two prototype smart homes, namely ARAS House A and ARAS House B. The homes are fully equipped with ambient sensors and are set up for experimental purpose. The annotation in ARAS is done by residents with the support of a software installation in their desktop. For activity recognition, studies in both ARAS and CASAS projects show that HMMs and CRFs have been the most successful techniques [15–17, 20, 21]. Recent work on combining dependencies can improve the recognition accuracy with a large margin [20]. For more detail, readers are encouraged to take a look at the survey [27]. Different from previous work, in MRAR system we employ deep learning techniques for activity recognition.

### 12.3 SSH: Smarter Safer Home Platform

Like most developed countries, Australia is now an ageing society as a result of sustained low fertility and increasing life expectancy. Population ageing is now having a range of implications for Australia, including soaring health expenditure, shortage of nursing and aged care staff and inadequate number of residential care placements with prohibitive entry fee. To help to fight these issues, the CSIRO research team has developed an innovative in-home monitoring and data analytics platform, the SSH platform, that seeks to support and extend independence and improve quality of life for aged residents through the use of cutting-edge pervasive communication and wireless sensor and monitoring technology. The potential benefits of these technologies are multiplied where distance separates families and adds substantial costs to delivery of health and other services.

The CSIRO SSH platform takes advantage of a provisional patent for a personalised measure of functional independence, indexed through the 'Objective Activity of Daily Livin' [49] which reflects their health and well-being status.

The platform includes a sensor-based in-home monitoring system (data collection), a cloud computing server (data analyses) and a client module (data presentation) with an iPad app, a family/carer portal and a clinical portal. The novelty of the SSH platform is its features of providing an objective and personalised measure of ADL components and scoring through non-wearable and non-intrusive sensors in the home environment and the ability to correlate this measure with self- or care-reported status of health and well-being.

### ***12.3.1 SSH Mobile Application***

To access the progress and summarised information derived from the SSH platform, residents are provided with an iPad and ‘Smarter Safer Homes’ app. The app interface was designed with prospective residents [43] and displays the progress status of their activities of daily living, vital signs and physical activity. It also provides personalised alerts and enables a real-time progress update, when deemed necessary by care provider(s). Residents can connect to their family or care services via a video conferencing services within the app.

An example of the app’s dashboard reflecting the daily status of health and well-being is represented by the different coloured rays. A three-quarter extension of the ray indicates individuals achieving their expected goal of health or well-being measures, whereas a ray below is a decline and full ray is an increased in their state of well-being. The design aims at friendly usability that makes the app more accessible to senior citizens.

### ***12.3.2 Family and Service Provider Portal***

Family members and friends of the elderly living alone often are anxious about their welfare. The platform includes a family portal that allows significant others to gain an insight into the lives of the elderly resident by communicating some of the information pertaining to their everyday lives via a website. There are four levels of access that the resident can make available to family members or nominated contacts.

In addition to the family portal, we also provide a web-based tool for experts. The clinical portal provides access for health professionals, such as nursing services or GPs, engaged to monitor the resident’s medical profile and health progress. The clinical portal has the capacity to present an individual’s health progress over various time periods (e.g. weekly or monthly). The portal can be accessed by multidisciplinary healthcare teams engaged in an individual’s care.

### 12.3.3 Smarter Safer Home Framework

The core principle of the SSH platform is to infer health and well-being status of residents through wireless sensor-based in-home monitoring system. Almost all sensors, except power sensor, are battery powered to increase installation flexibility and decrease maintenance requirements. Each of the sensors gathers data about a different aspect of resident lifestyle and activities contributing to functional independence. In this system, we use multi-sensor devices which can detect multiple changes from motion, temperature, light, humidity and vibration. The multi-sensor is small and light, runs on battery power and can be easily mounted to any surface. In this smart home, sensors are mounted in every room usually with daily access, such as bedroom, bathroom, kitchen, lounge and study room. Table 12.1 illustrates an example of some sensor data collected.

**Motion:** Motion sensors use PIR (pyroelectric infrared) to detect whether a person has moved in or out of the sensors range. PIR sensors work by comparing ambient background infrared compared with a current measured level of infrared. Using this technique has the advantage of high accuracy, low power consumption (allowing the sensors to last up to 2 years), as well as a long range (5 m) and wide beam width ( $120^\circ$ ).

**Temperature:** Temperature sensors use a thermistor to measure and report the air temperature. Thermistors are special resistors, which have a resistance that is heavily dependent on temperature. By using a model, the resistance of this thermistor can directly be related to temperature within the range of  $-10$  to  $50^\circ\text{C}$ .

**Light:** Light sensors use phototransistors to measure and report the ambient light level of the air, measured in Lux. A phototransistor is a light-sensitive transistor, which limits the amount of current let through in a circuit based on the level of light. The brighter the light level in the room, the more current passes through the circuit. This can be related to the Lux, which is the SI unit of luminous flux per unit area. 0 lx represents total darkness, 300–500 lx normal indoor lighting and 10,000+ lx for full outdoor sunlight

**Power:** The power sensors use an AC wall adapter to measure the power usage of any electrical device plugged into them. Measured in watts, these sensors are used

**Table 12.1** Example of sensors' data collected by the SSH

| ID  | Type               | Timestamp           | Values                |
|-----|--------------------|---------------------|-----------------------|
| 17  | Motion sensor      | 2018-09-07T06:24:00 | 1                     |
| 31  | Power sensor       | 2018-09-07T07:05:00 | 0.239 kw              |
| 148 | Humidity sensor    | 2018-09-07T08:15:00 | 58%                   |
| 143 | Temperature sensor | 2018-09-07T08:15:00 | 15.6 $^\circ\text{C}$ |

to determine when a device is being used and for what purpose this device is used for.

**Humidity:** Humidity sensors measure and report the ratio of moisture in the air. Humidity sensors use small capacitors whose capacitance will change with the absence of moisture in the air.

## 12.4 Multi-resident Activity Monitoring with SSH

### 12.4.1 *Speech Annotation*

In our system, we provide another way to annotate activities which use verbal input from residents instead of text input. This is natural for residents to speak with their smart home to provide information about activities in the form of languages. Our system takes the text format of the information and applies natural language processing to obtain ground truth for our recognition module. The speech-based annotation module in our system consists of a speech logger to record users' verbal inputs and a rule-based algorithm to extract the knowledge about identities, locations, times and activities, if any of those is available.

#### 12.4.1.1 **Speech Loggers**

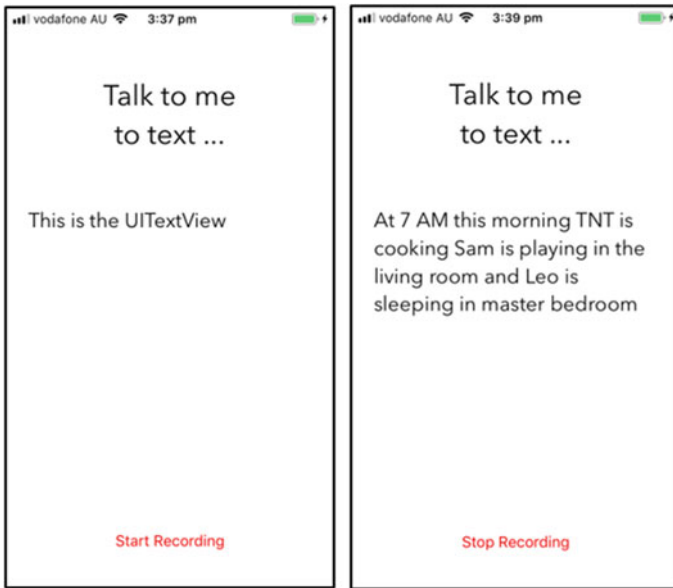
**iOS App:** We developed an iOS application to record the input from users. It allows them to make a report any time at their convenience. The voice signal is converted into text using speech-to-text API provided by Google. The texts then are sent to a secured Google drive account from which they are collected for the next step of knowledge extraction. The app simply has a button to let users to start and stop recording, as shown in Fig. 12.2 in next page.

The advantage of the mobile app is that it can be used at any location in the house. Annotation time is also quick as users only need to touch the screen twice.

**Google Home:** In some cases, especially where our residents are senior citizens or people with difficulty, the mobile app seems not the best choice. Therefore, we provide an alternative for speech logging using virtual assistants.

Virtual assistants are software agents which are integrated into smart devices which processes user input to perform tasks. In the last decade, virtual assistants have become mainstream with every major smartphone manufacturer developing their own agents. The most common are Siri for Apple, Alexa for Amazon, Cortana for Microsoft and the Google Assistant. Assistants can be used for a wide range of actions, such as general conversation, smart home control, information querying, media streaming, phone calls and schedule planning. The assistants receive commands through either text or speech and use NLP to analyse this input and execute a corresponding task. These assistants will also often use machine learning techniques





**Fig. 12.2** iOS Voice Logger App

to distinguish the voices from multiple users and to develop more accurate responses based on previous queries. Certain assistants such as the Alexa and Google Assistant make use of smart speakers, which are voice controlled wireless microphone/speaker devices. The advantage of smart speakers is that they are hands-free and can be integrated into a permanent home setup easily. These smart speakers listen passively for a designated keyword that triggers the built-in assistant. With these features, virtual assistants in smart speakers become ideal for multi residential homes. To obtain ground truth for our installed smart home, a smart speaker was used as an audio diary for the residents. The smart speaker was placed in a high-traffic area, and the residents verbally logged their daily activities. This data was then wirelessly uploaded to a database with a corresponding log time.

### **12.4.2 Knowledge Extraction**

After having the logged text, we apply part-of-speech tagging (POS) to obtain the structure of each sentence. The first group part-of-speech tags that we are interested in are the NN (noun, singular or mass), NNS (noun, plural), NNP (proper noun, singular) and NNPS (proper noun, plural) for the resident identities. The second group of tags we want to extract are: VB (verb, base form), VBD (verb, past tense), VBG (verb, gerund or present participle), VBN (verb, past participle), VBP (verb, non-3rd person singular present) and VBZ (verb, 3rd person singular present) which have

information about the activities. It is worth noting that we do not require residents to always log the activities at the time of recording. They can do it at any time that is convenient to them, i.e. logging the activities that they have done in the past along with the time. For that, we also extract cardinal numbers from CD (cardinal number) tag.

With the name of the residents, it is difficult for a computer program to identify different speech-to-text conversions due to the imperfection of the software. For example, one resident is logged by different ‘names’ in our system such as: Lin, Lynn and Lyn. Therefore, we need some human intervention to solve this issue.

Similarly, there might exist some different words for same activities, i.e. ‘having dinner’, ‘eat supper’ and ‘have meal at 7 PM’. More importantly, various expressions of same activities may be provided, i.e. sometimes it is a verb such as ‘prepare’ and ‘walk’, and sometimes it is a noun such as ‘preparation’ and ‘walking’. In order to get better annotation, we apply stemming and lemmatisation to group similar words together. We also create a set of rules to map different phrases/words to a set of activities.

### 12.4.3 Recognition Module

In this module we use multi-label recurrent neural networks (RNNs) for modelling activities of multiple residents. Previous works employ CRFs and HMMs which produce good results [17, 19, 20, 33]. We will show in the experiments that multi-label RNN is better than the state of the art and therefore has been chosen for our recognition modules.

The multi-label RNN is a RNN with multiple output layers which share the same hidden layer. Each output layer  $a^m$  is a softmax group of units, representing the distribution of the activities of resident  $m$ . Suppose that the set of activities of resident  $m$  is  $K^m$ , then the probability of an activity of this resident at time  $t$  is:

$$p(a^{m,t} = k^m) = \frac{\exp(\mathbf{h}^t{}^T \mathbf{u}_{k^m}^m)}{\sum_{k' \in K^m} \exp(\mathbf{h}^t{}^T \mathbf{u}_{k'}^m)} \quad (12.1)$$

where  $\mathbf{u}_k^m$  is a column vector of the weight matrix  $U^m$  connecting the hidden units to the units of softmax group  $m$ ;  $\mathbf{h}^m$  is the state of hidden layer at time  $t$ , which is calculated as below in the case of gated recurrent units [37].

$$\mathbf{z}^t = \text{sigmoid}(W_z^1 \mathbf{x}^t + W_z^2 \mathbf{x}^{t-1} + \mathbf{b}_z) \quad (12.2)$$

$$\mathbf{r}^t = \text{sigmoid}(W_r^1 \mathbf{x}^t + W_r^2 \mathbf{x}^{t-1} + \mathbf{b}_r) \quad (12.3)$$

$$\tilde{\mathbf{h}}^t = \text{sigmoid}(W^1 \mathbf{x}^t + W^2(\mathbf{r}^t \odot \mathbf{h}^{t-1})) \quad (12.4)$$

$$h^t = (1 - z^t) \odot h^{t-1} + z^t \odot \tilde{h}^t \tag{12.5}$$

where  $x^t$  is the input vector,  $z^t$  is the update gate vector, and  $r^t$  is the reset gate vector. Here,  $\odot$  denotes the element-wise matrix multiplication.

The advantage of recurrent neural networks is that it can perform online learning using stochastic gradient descent. This means that anytime the system achieves a logging from the residents we can update the recognition model on the fly.

## 12.5 Deployment and Evaluation

### 12.5.1 Set-up

We set up the MRAR system at a house lived by two residents and a kid. The layout of the house is shown in Fig. 12.3a.

For collecting the data, we use multi-sensors which are the combination of five sensors: motion, temperature, light, humidity and vibration, as in Fig. 12.3c. The multi-sensor does not interfere with the residents' daily living as it requires infrequent battery changes and passively collects data without intervention. The main features of the data collected will be determining when someone enters/exits a room and any activities that cause changes in temperature, light or humidity such as taking a shower. This data is sent to the gateway (see Fig. 12.3e), which is placed in a central location at the home and recorded in a database. The power sensors as shown in Fig. 12.3b

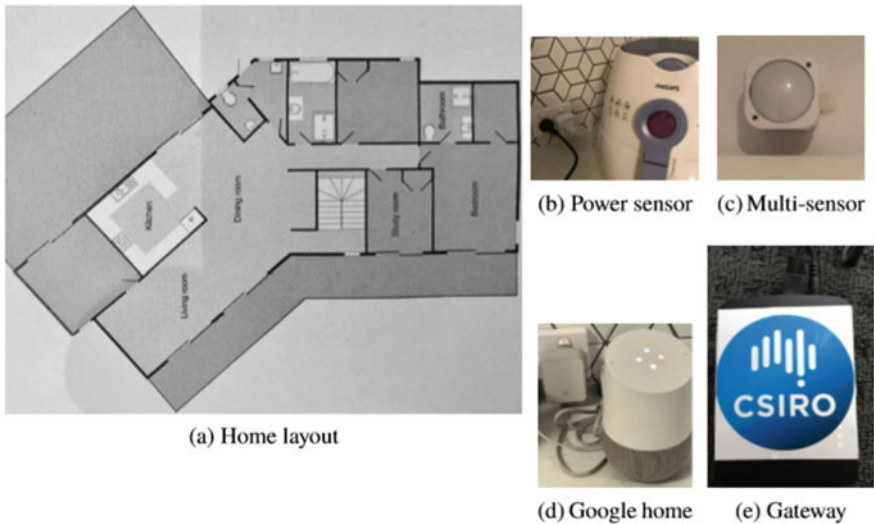


Fig. 12.3 Home layout, sensors and gateway

were installed in the kitchen connected to the most used kitchen appliances. These sensors report back power usage of the sensor constantly, and this data is used to determine the time and duration an appliance is used. We install 8 multisensory devices and 3 power sensors to make a total of 27 sensors in the house. The data is sent to the gateway and recorded in the database and is used to determine when meal preparation is occurring.

We also set up a recording device in kitchen where anyone can verbally log the activities of everyone in the house at any time that is convenient for them, see Fig. 12.3d. At this stage, the main purpose of the recording device is for annotating the activities. Some devices such as Google home and Alexa can provide speech-to-text data which is easier for processing. Note that residents always have an alternative to log their activities by using the iOS app as mentioned earlier in Sect. 12.4.1.1. Samples of recordings in text format are shown in Table 12.2. As we can see, converting speech to text is not always perfect since errors can come from accents, mispronunciation, interference of background noise, etc. Therefore, we use NLP techniques to minimise the effect of such disadvantage and also to identify the activity of a resident at each recording time. It is worth noting that, different from the other works, we do not label all activities in a day. Instead, we do weakly labelling at several times of a day at residents' convenience. This is repeated for several days to cover all possible regularities. In this experiment, we perform verbal logging for 10 days at the same time we collect the data from our sensors.

**Table 12.2** Samples of recordings

| Date time              | Recording  |
|------------------------|--|
| Thu Sep 28, 2017 22:05 | 5–10 R1 taking shower                                    |
| Thu Sep 28, 2017 22:04 | 955 R1 taking shower                                     |
| Thu Sep 28, 2017 21:17 | 9:15 R3 play Caught in the living room                   |
| Thu Sep 28, 2017 21:17 | Call across my media playing Gang in the meeting         |
| Thu Sep 28, 2017 21:16 | R1 caught a possum oh R3 play calling in the living room |
| Thu Sep 28, 2017 21:16 | 915 LED R3 upgrade toy in living room                    |
| Thu Sep 28, 2017 21:16 | May 15th R3 play for in the living room                  |

Residents' real names are replaced by R1, R2 and R3

### 12.5.2 Annotation

Applying the NLP techniques mentioned in Sect. 12.4.2, we are able to extract the activities in this smart home. Here, we show several samples of words after applying stemming and lemmatisation as below:

| VBG             |             |          |          |          |           |           |
|-----------------|-------------|----------|----------|----------|-----------|-----------|
| Before stemming | 'preparing' | 'having' | 'making' | 'taking' | 'playing' | 'calling' |
| After stemming  | 'prepare'   | 'have'   | 'make'   | 'take'   | 'play'    | 'call'    |

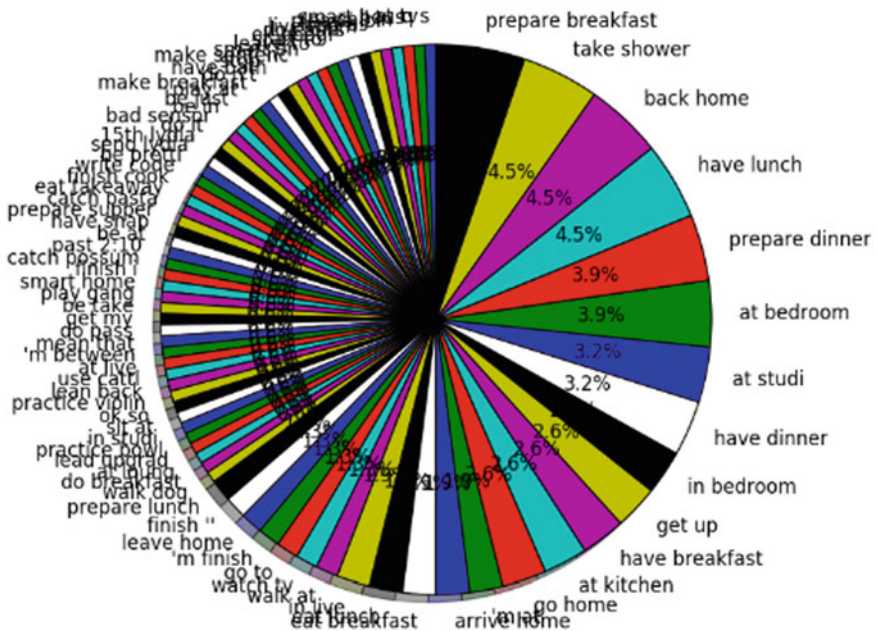
| VBD                  |         |        |          |            |        |       |        |
|----------------------|---------|--------|----------|------------|--------|-------|--------|
| Before lemmatisation | 'left'  | 'went' | 'caught' | 'finished' | 'took' | 'got' | 'had'  |
| After lemmatisation  | 'leave' | 'go'   | 'catch'  | 'finish'   | 'take' | 'get' | 'have' |

The date and time (also sometimes other numbers) are partitioned in **CD** tag as: [‘8:30’, ‘7:30’, ‘9’, ‘738’, ‘7:45’, ‘6:45’, ‘8’, ‘7:15’, ‘7’, ‘6:15’, ‘9:30’, ‘3’, ‘8:45’, ‘10:15’, ‘1025’, ‘10:25’, ‘955’, ‘9:15’, ‘915’, ‘15th’, ‘7am’ ...]

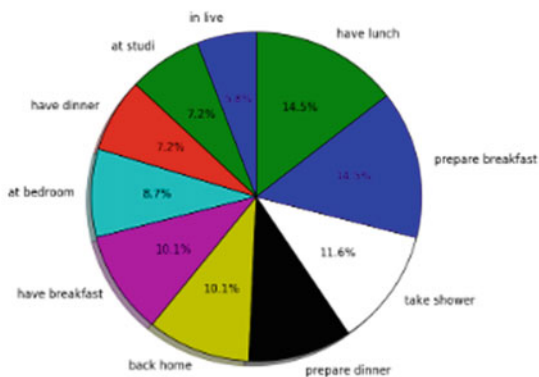
In Fig. 12.4a, we show all ‘activities’ extracted from the logged speech-to-text data. As we can see, there is some noise in this extraction due to the logging error. However, it is trivial and we can fine-tune the process to get the activities of interests.

### 12.5.3 Activity Recognition

We ordered all activities and selected 10 most occurred ones for prediction as shown in Fig. 12.4b. Due to the sparse of the logging at this preliminary stage, we group all activities in living room into one activity, as illustrated as ‘in live’. Similarly, we also group activities in study room for ‘at study’ and activities in bedroom for ‘at bedroom’. In this experiment, we use recurrent neural networks with a sharing hidden layer and multiple output layers, and each present activities of a resident. Before applying the multilab RNNs to our system, we perform a test on three smart homes from two benchmarks projects CASAS [33, 50] and ARAS [15]. We compare the model with popular models in the field such as factorial hidden Markov models [20], factorial random field [51] and incremental decision tree [52]. In this case, we use RNN with recurrent gated units [37]. In CASAS data, we use the data of 24 days for training, 1 day for validation and 1 day for testing. In ARAS, the data of both



(a) All “extracted activities.”



(b) 10 most occurred activities.

Fig. 12.4 Extraction of activities from text

**Table 12.3** Performance of multi-label RNN compared to other methods

|             | CASAS (%)        | ARAS House A (%) | ARAS House B (%) |
|-------------|------------------|------------------|------------------|
| fHMM        | 55.43            | 39.33            | 78.91            |
| fCRF        | 45.84            | 55.95            | 74.44            |
| iDT         | N/A              | 48.36            | 64.19            |
| multilabRNN | 80.63 $\pm$ 3.26 | 55.78 $\pm$ 0.29 | 76.93 $\pm$ 0.20 |

House A and House B are partitioned into 7 days for training, 2 days for validation and 2 days for testing.

For the multi-label RNN, we search for the hyper-parameters: learning rate and hidden layer size using grid-like search. We search the learning rate in {0.0001, 0.001, 0.01, 0.1, 0.3, 0.5} and the hidden layer size in {10, 50, 100, 500}. The results show that multilab RNN outperforms factorial hidden Markov models (fHMMs), factorial conditional random field (fCRF) and incremental decision tree (iDT) in CASAS data. In particular, we can observe at least 25.2%. In ARAS House A, multi-label RNN achieves comparable accuracy to factorial conditional random field which are both better than factorial hidden Markov model and incremental decision tree. In ARAS House B, our model's performance is slightly lower than the factorial hidden Markov model, but it is still higher than the rest. Therefore, we choose multilab RNN for our system. We use activity in 6 days to train the RNNs and test the model in 4 days which achieve 72.21% accuracy (Table 12.3).

## 12.6 Conclusions and Future Work

We have shown a multi-activity recognition system for ambient smart homes. Our system is capable to transform a normal house to a smart home with minimum effect. In particular, we do not require residents to be visually monitored or to wear any electronic device. We also provide an annotation tool that does not require residents to work with smartphones or computers. Instead, they can talk with the system to input the ground truths at their convenience. This reduces the manual effort of annotation experts considerably.

The major disadvantage of such system is that our voice loggers are only be able to passively record the voice messages triggered by residents. In some cases, this led to very sparse labelled data which is not useful for training as our RNNs use sequential samples as inputs. Another issue is that our labelling technique depends heavily on the quality of the speech-to-text API provided by Google, which sometimes struggles to understand the context of conversations and provides misleading outputs. For the future work, we will work on the development of a smart logger that is able to actively communicate with residents to ask for activity based on the need of our system while, at the same time, can also act as a virtual companion. We will also improve the labelling approach to make the system more accurate.

## References

1. Jiang L, Liu D-Y, Yang B (2004) Smart home research. In: Proceedings of International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826), vol 2, pp 659–663
2. Lutolf R (1992) Smart home concept and the integration of energy meters into a home based system. In: Proceedings of the 7th International Conference on Metering Apparatus and Tariffs for Electricity Supply, pp 277–278
3. Cook DJ, Das SK (2007) How smart are our environments? an updated look at the state of the art. *Pervasive Mob Comput* 3(2):53–73
4. Cucchiara R, Grana C, Prati A, Vezzani R (2005) Computer vision system for in-house video surveillance. In: IEEE Proceedings-Vision, Image, and Signal Processing, 2005, pp 242–249
5. Nguyen N, Venkatesh S, Bui H (2006) Recognising behaviours of multiple people with hierarchical probabilistic model and statistical data association. In: Proceedings of the 17th British Machine Vision Conference (BMVC'06). The British Machine Vision Association and Society for Pattern Recognition, pp 1239–1248
6. McCowan I, Gatica-Perez D, Bengio S, Lathoud G, Barnard M, Zhang D (2005) Automatic analysis of multimodal group actions in meetings. *IEEE Trans Pattern Anal Mach Intell* 27(3):305–317. <https://doi.org/10.1109/TPAMI.2005.49>
7. Du Y, Chen F, Xu W (2007) Human interaction representation and recognition through motion decomposition. *IEEE Signal Process Lett* 14(12):952–955. <https://doi.org/10.1109/lsp.2007.908035>
8. Du Y, Chen F, Xu W, Li Y (2006) Recognizing interaction activities using dynamic Bayesian network. In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), pp 618–621
9. Jalal A, Uddin MZ, Kim T-S (2012) Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Trans Consum Electron* 58(3):863–871
10. Han J, Pauwels EJ, de Zeeuw PM, de With PHN (2012) Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. *IEEE Trans Consum Electron* 58(2):255–263
11. Yamamoto Y, Yoda I, Sakaue K (2004) Arm-pointing gesture interface using surrounded stereo cameras system. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, vol 4. Cambridge, pp 965–970
12. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum D (2016) Recognizing complex activities by a probabilistic interval-based model. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, pp 1266–1272
13. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2activity: recognizing complex activities from sensor data. In Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15. AAAI Press, pp 1617–1623
14. Plötz T, Hammerla NY, Olivier P (2011) Feature learning for activity recognition in ubiquitous computing. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI'11. AAAI Press, pp 1729–1734
15. Alemdar H, Ertan H, Incel OD, Ersoy C (2013) Aras human activity datasets in multiple homes with multiple residents. In: Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth '13, ICST, Brussels, Belgium, pp 232–235
16. Chen R, Tong Y (2014) A two-stage method for solving multi-resident activity recognition in smart environments. *Entropy* 16(4):2184
17. Chiang Y-T, Hsu KC, Lu CH, Fu L-C, Hsu JY-J (2010) Interaction models for multiple-resident activity recognition in a smart home. In: IEEE/RSJ International Conference on IROS, pp 3753–3758
18. Cook DJ (2012) Learning setting-generalized activity models for smart spaces. *IEEE Intell Syst* 27:32–38



19. Singla G, Cook DJ, Schmitter-Edgecombe M (2010) Recognizing independent and joint activities among multiple residents in smart environments. *J Ambient Intell Humaniz Comput* 1(1):57–63
20. Tran SN, Zhang Q, Karunanithi M (2017) Improving multi-resident activity recognition for smarter homes. In: *IJCAI WS on AI for IoT*
21. Wang L, Tao G, Tao X, Chen H, Jian L (2011) Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive Mob Comput* 7(3):287–298
22. Khan SS, Karg ME, Hoey J, Kulic D (2012) Towards the detection of unusual temporal events during activities using HMMs. In: *Proceedings of the Conference on Ubiquitous Computing (UbiComp'12)*. ACM, New York, pp 1075–1084. <https://doi.org/10.1145/2370216.2370444>
23. Riboni D, Pareschi L, Radaelli L, Bettini C (2011) Is ontology-based activity recognition really effective? In: *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops'11)*, pp 427–431. <https://doi.org/10.1109/percomw.2011.5766927>
24. van Kasteren TLM, Englebienne G, Kröse BJA (2011) Hierarchical activity recognition using automatically clustered actions. In *Proceedings of the 2nd International Conference on Ambient Intelligence (AmI'11)*, pp 82–91
25. van Kasteren TLM, Englebienne G, Krose BJA (2010) Human activity recognition from wireless sensor network data: benchmark and software. In: Chen L, Nugent CD, Biswas J, Hoey J (eds) *Activity recognition in pervasive intelligent environments*. Atlantis ambient and pervasive intelligence. Atlantis Press, pp 165–186
26. van Kasteren TLM, Noulas A, Englebienne G, Krose B (2008) Accurate activity recognition in a home setting. In: *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp'08)*. ACM, pp 1–9. <https://doi.org/10.1145/1409635.1409637>
27. Benmansour A, Bouchachia A, Feham M (2015) Multioccupant activity recognition in pervasive smart home environments. *ACM Comput Surv* 48(3):34:1–34:36
28. Liao L, Fox D, Kautz H (2005) Location-based activity recognition using relational Markov networks. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, pp 773–778
29. Munguia-Tapia E, Intille SS, Larson K (2004) Activity recognition in the home using simple and ubiquitous sensors. In: *Proceedings of Pervasive*, pp 158–175
30. Philipose M, Fishkin K, Perkowski M, Patterson D, Fox D, Kautz H, Hahnel D (2004) Inferring activities from interactions with objects. *IEEE Pervasive Comput* 3:50–57
31. Szewczyk S, Dwan K, Minor B, Swelove B, Cook D (2009) Annotating smart environment sensor data for activity learning. In: *Technology and Health Care, special issue on Smart Environments: Technology to support health care*
32. Chen C, Cook D (2012) Behavior-based home energy prediction. In: *Proceedings of the International Conference on Intelligent Environments*
33. Hsu K-C, Chiang Y-T, Lin G-Y, Lu C-H, Hsu JY-J, Fu L-C (2010) Strategies for inference mechanism of conditional random fields for multiple-resident activity recognition in a smart home. In: *Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, Berlin, pp 417–426
34. Crandall AS, Cook DJ (2008) Resident and caregiver: handling multiple people in a smart carefacility. In: *Proceedings of the AAAI Fall Symposium—AI in Eldercare: New Solutions to Old Problems*, pp 39–47. <http://dx.doi.org/10.1.1.329.4737>
35. Crandall AS, Cook DJ (2008) Attributing events to individuals in multi-inhabitant environments. In *Proceedings of the IET 4th International Conference on Intelligent Environments (August 2008)*, pp 1–8. <https://doi.org/10.1049/cp:20081164>
36. Tunca C, Alemdar H, Ertan H, Incel OD, Ersoy C (2014) Multimodal wireless sensor network-based ambient assisted living in real homes with multiple residents. *Sensors*
37. Chung J, Gülçehre Ç, Cho KH, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555
38. Just Checking (2016) <http://www.justchecking.com.au>
39. My Lively (2016) <http://www.mylively.com>

40. Tunstall (2012) Solution sheet ADLife an activities of daily living (ADL) monitoring. <http://www.tunstall.co.uk/>
41. Telemedcare (2016) <http://telemedcare.com>
42. Care Innovations (2016) <http://careinnovations.com>
43. Bradford D, Freyne J, Karunanithi M (2013) Sensors on my bed: the ups and downs of in-home monitoring. In: Inclusive Society: Health and Well-being in the Community, and Care at Home. Proceedings of the 11th International Conference on Smart Homes and Health Telematics, ICOST 2013, Singapore, pp 10–18
44. Chen L, Nugent CD, Wang H (2012) A knowledge-driven approach to activity recognition in smart homes. *IEEE Trans Knowl Data Eng* 24(6):961–974
45. Nugent CD, Chen L, Okeyo G (2014) An ontology-based hybrid approach to activity modelling for smart homes. *IEEE Trans Hum-Mach Syst* 44(1):92–105
46. Alshammari NO et al (2017) OpenSHS: open smart home simulator. *Sensors*
47. WSU (2007) Washington State University Casas project. <http://casas.wsu.edu>
48. Feuz KD, Cook DJ (2013) Real-time annotation tool (RAT). In: Proceedings of the AAAI Workshop on Activity Context-Aware System Architectures
49. Karunanithi M, Zhang Q (2015) Objective assessment and scoring of activities of daily living. Pending Patent
50. Cook DJ, Crandall A, Singla G, Thomas B (2010) Detection of social interaction in smart spaces. *Cybern Syst* 41(2):90–104
51. Sutton C, McCallum A, Rohanimanesh K (2007) Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data. *J Mach Learn Res* 8:693–723
52. Prosegger M, Bouchachia A (2014) Multi-resident activity recognition using incremental decision trees. In: Adaptive and Intelligent Systems—Third International Conference, ICAIS 2014, Bournemouth, UK, pp 182–191

# Chapter 13

## New Environments for the Evaluation of Smart Living Solutions



**Beatriz Merino Barbancho, Ivana Lombroni, Cecilia Vera-Muñoz and María Teresa Arredondo**

**Abstract** In recent years, the evolution of the technology and its adoption by the citizens has made possible the inclusion of new tools in the innovation processes. These tools have allowed the development of better products and services and the reduction of the time to market of these solutions for the benefit of all end users. In this context, the concept of Living Lab has emerged, as an open innovation ecosystem where stakeholders, including professionals, developers and end users, can cooperate on solutions to address relevant challenges using co-creation and evaluation methodologies. Living Labs can support and enhance the innovation process throughout the different phases of the value chain, and they can also act as the connectors between the needs (the users) and the offer (the suppliers). This chapter presents the Living Labs as the novel instruments for evaluating, assessing and validating innovative products, solutions or services in the particular domain of smart living environments.

**Keywords** Living labs · Innovation · Validation and evaluation · Market access

### 13.1 Introduction

The advances that the new technologies have experienced during the last decades have enabled the implementation of new products, services and solutions that are improving every single aspect of our daily lives. The development of products focused on solving health problems or aiming at enhancing our daily living environments has not been an exception in this tendency, and new solutions are appearing each year trying to make our lives easier or to solve some particular health-related problems. As a result, our living environments are progressively equipped with a set of new products and technologies that we can name smart living solutions.

However, despite the increasing amount of new available smart living products nowadays, many developments still remain at a research level and never reach the

---

B. Merino Barbancho (✉) · I. Lombroni · C. Vera-Muñoz · M. T. Arredondo  
Life Supporting Technologies, Universidad Politécnica de Madrid, Madrid, Spain  
e-mail: [bmerino@lst.tfo.upm.es](mailto:bmerino@lst.tfo.upm.es)

E.T.S.I. Telecomunicación, Avenida Complutense 30, 28040 Madrid, Spain

© Springer Nature Switzerland AG 2020  
F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications  
and Networks, [https://doi.org/10.1007/978-3-030-25590-9\\_13](https://doi.org/10.1007/978-3-030-25590-9_13)

market. Moreover, from those products that are actually commercialized, only a very small percentage is actually adopted by the population and integrated into their daily lives in the long term. The reasons for this are usually a wrong approach in the design of the product or the solution, or in its validation process.

A couple of decades ago, back in the 2000s, a new concept started to emerge: the concept of Living Lab. The Living Labs promised to be the link between research and the real world, as an enabler to enhance the outcomes in such a way that were more adapted to the end users. These organizations became very popular during the first decade of the century. A significant number of Living Labs were created at that time all around the world and particularly in Europe, with the establishment of the European Network of Living Labs (ENoLL) [8].

Although the popularity and utilization of Living Labs decreased after a few years since their creation, there has been an upturn on the use of this type of organizations lately. The main causes for this have been the following: (a) first, the change of focus in the Europe from research to innovation, and the increasing interests, at all levels, to transfer innovations to the market; and (b) secondly, the need to perform an increasing number of evaluations on new products and solutions, in order to cope with the new regulations, and to facilitate the adoption of innovations into the market.

In addition, some organizations, like the EIT Health [7], have detected the potential that a new use of Living Labs could have in the innovation domain, and the added value of utilizing these facilities to help to get products into the market with an increased probability of success. This has also helped to put Living Labs in the spotlight again. In this context, the Living Labs organization have re-emerged and have started to adapt their activities to cover these needs.

In this chapter, we present a new vision of the Living Labs, as novel instruments for evaluating, assessing and validating innovative products, solutions or services in the particular domain of smart living environments.

## 13.2 Living Labs: The Concept

The concept of Living Lab, as well as its real origins, is not clearly defined nor universally accepted. Many authors consider that William J. Mitchell, from the Massachusetts Institute of Technology (MIT), was the creator of the concept of Living Lab. He proposed, in the early 2000s, to start testing some of the developments made at research level in real buildings or controlled areas in specific neighbourhoods [19]. However, some authors like Van Geenhuizen M. et al. or Bajgier S. M. situate their origin even earlier, back to the early 1990s, when “some restricted city neighbourhoods were used to teach students to solve real-world problems with the help of relevant stakeholders” [3, 25].

The truth is that the concept of Living Lab started to emerge and become known in the early 2000s, with the appearance of the Ambient Assisted Living paradigm and the adoption of new techniques in the research domain, like the involvement of end users in the design phases. The initial Living Labs were focused on research and

provided a context where both researchers and end users could work together in the development and validation of different solutions.

In Europe, this concept gained a significant popularity with the creation of the European Network of Living Labs (ENoLL) in 2006, initially sponsored by the European Commission [8]. During the first years of its existence, this network managed to promote the concept and to support the creation of new Living Labs all across Europe. In addition, it gathered a relevant group of European Living Labs working in different fields. This European Network of Living Labs is still operative today, and it is an important reference in the European landscape. At the beginning of 2019, the network included almost 400 recognized Living Labs [8], which formed a heterogeneous group including organizations with very different approaches, activities and thematic specialization.

In recent years, some new networks of Living Labs have been created in Europe, with the aim of being more specialized and grouping mainly a particular type of Living Labs. This is the case of the EIT Health Living Labs network, created in 2016 under the framework of the EIT Health [7]. The EIT Health is an organization, supported by the European Union and its European Institute of Innovation and Technology (EIT) that aims to “gather the best-in-class health innovators to deliver solutions to enable European citizens to live longer, healthier lives by promoting innovation” [7]. The EIT Health Living Labs network was created with the objective to contribute to this general goal, by using Living Labs as a novel method to enhance the innovation process in the health and healthcare domains.

In addition, there have been also other initiatives, more at a local level, that have managed to create smaller but local specialized networks of Living Labs. A good example of this is the Forum LLSA (Le Forum des Living Labs en Santé et Autonomie), which gathers a relevant number of Living Labs in France specialized in health and autonomy [11].

### ***13.2.1 Definition of Living Lab***

Nowadays, almost three decades after the appearance of the initial Living Labs, there is still no universal or global recognized definition for this type of organizations. The concept of Living Lab has evolved along the years, and all these networks and Living Labs themselves had also to adapt themselves to new contexts and situations in order to survive.

The European Network of Living Labs (ENoLL) has always provided a definition for a Living Lab since its origin, and nowadays, it defines Living Labs as “user-centred, open innovation ecosystems based on systematic user co-creation approach, integrating research and innovation processes in real-life communities and settings” [8]. They also indicate that “Living Labs are both practice-driven organizations that facilitate and foster open, collaborative innovation and real-life environments or arenas where both open innovation and user innovation processes can be studied and subject to experiments and where new solutions are developed” [8]. Moreover,

according to them, “Living Labs operate as intermediaries among citizens, research organizations, companies, cities and regions for joint value co-creation, rapid prototyping or validation to scale up innovation and businesses. Living Labs have common elements but multiple different implementations” [8].

ENoLL has not been the only organization to provide a definition of the concept of Living Lab. According to Picard R. et al., one of the co-founders of the Forum LLSA, a Living Lab, is “a consultation mechanism bringing together public and private stakeholders, actors or professionals, to enrich the co-creation and evaluation process that offers a real-time ecosystem for experimentation with the purpose of promoting and validating different innovation projects” [11, 21]. Also, within the EIT Health Living Labs network, Living Labs are defined as “facilities dedicated to end user mediated business creation” [7]. According to this, the EIT Health considers that “Living Labs can help to enhance conception of fully market-ready products with high added value and high probability of success, throughout the different phases of the value chain” [7].

Some individual Living Labs have also provided their own definition of this concept. Insightfully, Bergvall-Kareborn et al. from the Halstad Living Lab [12] defined a Living Lab as “a set of public–private partnerships in which researchers, citizens, professionals, companies or government work together to create and validate new business ideas, services or technologies in a real environment” [4].

According to this definition, “the main objective of a Living Lab is, therefore, to have a shared space in which new forms of work can be developed with the end users in such a way that research and development is stimulated by being, citizens and users, the key participants in the process of innovation” [4]. In addition, this definition considers that there are several elements that need to be part of a Living Lab, and these include several actors (partner and users), as well some components like research, management and infrastructure, which are considered to be essential to accomplish a successful evaluation in these environments [12].

Despite the different approaches provided in all these definitions, there are several common and key elements in all of them that make Living Labs unique [9]. These include the use of co-creation methods, an active involvement of end users in the whole innovation process, the access to real-life (or almost real-life) environments, the utilization of specific methodologies to obtain evidence-based results, and the participation of multiple stakeholders.

All these aspects make these organizations perfect to perform evaluations of the new complex solutions that are now being developed, particularly in the smart living solutions domain.

### ***13.2.2 Classification Criteria and Relevant Dimensions in Living Labs***

As it occurs with the definition of Living Labs, there is not a common agreement on the relevant criteria to be used to qualify and classify these organizations.

The European Network of Living Labs (ENoLL) [10] made one of the first attempts to define these criteria and, after more than 20 years of existence, they have now a consolidated set of 20 criteria grouped in 6 main areas: active user involvement, co-creation, orchestration, multi-method approach, real-life setting, and multi-stakeholder participation [9]. A few years ago, the Forum LLSA developed a detailed self-assessment framework that included 59 criteria grouped in 11 dimensions: governance, strategy and policies, human resources, financial and technical resources, processes, research, human resources outcomes, client or user outcomes, societal outcomes, operational outcomes, and research outcomes [11].

On the other hand, the EIT Health Living Labs network recently built its own set of classification criteria and relevant dimensions. These were defined following an iterative process where co-creation activities had a fundamental role. In this process, the existing work already done by both ENoLL and the Forum LLSA was considered as a starting point, together with the input of experts from different Living Labs in Europe. As a result, the EIT Health Living Labs network is using a final set of nineteen criteria, grouped in seven relevant categories, that each of the organizations should meet in order to be part of the EIT Health Living Labs network. The seven categories cover the following aspects: speciality area, ecosystem, users, resources, methodology, business and track record. Table 13.1 summarizes the list of criteria under each category.

In addition to the definition of its own set of classification criteria, the EIT Health Living Labs network has also identified a set of 9 relevant dimensions for Living Labs, which are used to assess the maturity levels and the quality of service that each Living Lab provides. These dimensions are the following: governance of the Living Lab, human resources and Living Lab team, strategy and value proposition of the Living Lab, financial sustainability, operations and processes, monitoring and quality management, user and stakeholder involvement, and methodology and tools.

The previous examples show how, despite the existing common ground between the set of categories, dimension and criteria defined by each network, the high specialization and diversity of Living Labs makes necessarily the customization of such aspects at a very low detail level. This means that the Living Labs belonging to each of these networks, are perfectly characterized and classified according to the final objectives of the network. From the potential client point of view, this facilitates the process of selection of a Living Lab and, as a result, a good performance in the evaluation process due to the personalization of needs and services provided by each ecosystem.

**Table 13.1** Classification criteria for the EIT Health Living Labs network

| Key aspects    | Subgroups   |
|----------------|---|
| Specialty area | Area of specialization<br>Maturity level                                  |
| Ecosystem      | Stakeholders<br>Type of service<br>Context<br>Link with other initiatives |
| Users          | Type of users<br>Number of users<br>End user selection criteria           |
| Resources      | Support technology<br>Operational readiness<br>Infrastructure or setting  |
| Business       | Market information available<br>Regulations and ethics<br>IPR principles  |
| Methodology    | Methodology   |
| Track record   | Previous activities<br>Evidence   |

### 13.3 Methodologies

One of the main characteristics of a Living Lab is the use of specific methodologies in their activities. As explained above, this aspect is considered one of the nine relevant dimensions in a Living Lab according to the EIT Health classification criteria. The use of adequate methodologies enables these organizations to obtain evidence-based results when they perform any type of evaluation, and also allows them to provide recommendations for adapting the design of each solution to the particular needs of its intended end users.

By nature, Living Labs are different from each other. Each of them is usually specialized on a particular domain and, even among the ones working in the same field, there are always differences in terms of the ecosystem, the stakeholders or the end users they have access to or works with. Also, the methodologies used in Living Labs can be very diverse and depended basically on the activities performed and services offered by each of them.

The enumeration of all the methodologies and method commonly used in Living Labs would result in an endless list. Not only there is a large amount of different methods and methodologies to cover each of the innovation phases there are also new ones appearing continuously. However, there are some basic aspects that all the Living Labs have in common, like the involvement of end users and the use of co-creation methods, and this determines somehow the use of a specific type of methodologies. This is the case of the user-centred design (UCD) methodologies, which are broadly used in the majority of the existing Living Labs in Europe when



providing their services, or the use of specific evaluation and co-creation methods, also very relevant in the Living Labs activities.

This subsection presents an overview of some of these methodologies, and the use that Living Labs make of them, especially for providing validation and co-creation services.

### ***13.3.1 User-Centred Design Methodologies***

User-centred design (UCD) lies on the simple fact of considering of the users at the centre of the design. This concept is not new, and the International Organization for Standardization included it in one of their approved standards already in 1999: the standard ISO 13407, focused on human-centred design processes for interactive systems [16]. This standard provided a framework that ensured that the needs of the users were at the core of the design process. The ISO 13407 standard was updated and renamed in 2010, as ISO 9241-210 (last reviewed and confirmed in 2015) [17], and will be substituted by the ISO/FDIS 9241-210 soon (under development on April 2019) [15].

Since the publication of the first ISO standard, back in 1999, numerous and different so-called UCD methodologies arose, all sharing the same principles [24]:

- The involvement of users in the design and development phases of a solution and following an iterative process.
- A design based upon an explicit understanding of users, tasks and environments; driven and refined by user-centred evaluation; and addressing the whole user experience.
- The involvement of multidisciplinary skills and perspectives in the design team.

The number of existing UCD methodologies is endless and new ones are continuously appearing. Although they all follow the general principles of UCD, each of them makes a special emphasis on a particular aspect of the whole design and implementation process. In this sense, some of them cover the whole innovation design and development process (e.g. goal-directed design) whilst others are more focused on supporting a particular stage like the design phase (e.g. design-driven innovation, data- or metrics-driven design) or the business and market aspects (e.g. lead-user design, Lean start-up).

By nature, when the first Living Labs were created, they soon adopted these methodologies as a basis for their activities. This has been maintained during the years and, nowadays, the utilization of UCD methodologies has become intrinsically indispensable for the Living Labs activities.

### 13.3.2 Evaluation Methodologies

Despite its name, UCD methodologies are not only focused on the design phase, but they usually cover all the innovation paths. However, it is worth mentioning a particular group of these methodologies that are used in the validation stage. These are the evaluation methodologies, a very important group for the Living Labs activities.

An evaluation methodology can be defined as a tool or method to determine the level of performance of any aspect related to a product, solution or service. The traditional assessment performed with these methods tackles different and relevant aspects of the products and can be the following:

- *Technical evaluation*, to determine that the product or solution works as it is intended to do from the technical point of view. Different technologies require different methods to assess these aspects.
- *User Experience (UX)*. The ISO 9241-110:2010 defines user experience as “a person’s perceptions and responses that result from the use and/or anticipated use of a product, system or service” [17].
- *Usability and satisfaction*. A usability evaluation assesses the easiness to use a system, product or solution and how easy or difficult it is for users that utilize it for the first time to accomplish a certain task [20, 24]. Usability evaluations are commonly performed together with a satisfaction assessment, which intends to determine the level of satisfaction of the users with a particular product.
- *Market and business*. These are the most novel methods added to the validation phase and they focus on assessing the adaptation of a specific product to a concrete market. They also analyse some business aspects like the business models or business plans.

Some particular Living Labs have even developed their own methodologies for performing this type of evaluations. More specifically, the company imec, very well known for its activities in the Living Labs domain, has recently developed a new tool called “Living Lab Innovatrix” (© imec), with the goal to help entrepreneurs and businesses to develop new products or services [14]. As it is defined in its website, “the tool helps them to innovate in a structured manner, with a focus on all types of users with their specific wishes and requirements” [14].

There are several techniques and methods that can be used to perform any of these evaluations, being the most commonly used focus groups, usability testing, card sorting, participatory design, questionnaires and interviews.

In principle, any organization with the right training and knowledge of a specific methodology could perform an evaluation of their solution, product or service. However, it is not very common to count with this particular expertise within the development teams. This is especially true in the case of SMEs or start-ups, which are usually focused more on the technical developments and business-related aspects than in the validation phase of their solutions.

The Living Labs can play an important role in this evaluation phase because, not only they are experts in applying different methodologies, they also have access

to a complete ecosystem (composed by end users and relevant stakeholders) that facilitates the evaluation processes and ensure high-quality results. Moreover, the Living Labs can also play an important role in the earliest stages of the development of a solution (since its conception).

### 13.4 Living Labs and Innovation

Living Labs, as a whole, constitute a complex and very heterogeneous group. The great diversity among Living Labs makes them a difficult group of organizations to catalogue, classify or work with. However, this variety also offers great opportunities for utilizing Living Labs in a large number of different contexts and, thus, benefiting from using this unique type of organizations.

The usual activities of initial Living Labs were focused on research and the organization of pilot testing activities with end users in real-life environments. Very soon, these activities evolved with the development of concepts like co-creation or open innovation that were incorporated into the normal activities performed in Living Labs. The added value of Living Labs increased significantly when they started to work on activities that were not only focused on research, but on supporting innovation processes.

Innovation can be defined as the process of transferring an idea into a product or service that creates value for the community. This process includes several phases that cover the whole innovation chain like the product conception, the product design, the product development, the product manufacturing and the marketing, among others [23]. In order to be effective and successful, this process cannot be linear, it has to be continuously iterative. This means that some aspects, like the definition of a concept, the matching of the market needs to technological capabilities or the performance of research activities to fill gaps in knowledge, have to be revisited several times along the way until the product is totally finalized and ready to go to market [23].

From this general term, the concept of open innovation is derived, as a new innovation strategy through which clients go beyond their limits and develop cooperation with external organizations or professionals [5].

The involvement of Living Labs in the innovation process can add a significant value. This is particularly relevant in the area of smart living solutions, because the increasing development of new products and services, many of them addressed to the citizens as end consumers, makes impossible to discern which are the most adequate for each particular case.

The smart living solutions require very careful validation processes according to the established rules, goals and standards. The assessment of their performance usually implies strong interactions with other existing provider solutions, as well as with different kinds and cohorts of users that need and request very specific experimental set-ups. The success of a new solution validation is highly dependent on a careful design of the whole evaluation process, in a way that ensures a high correspondence between the solution to be assessed and the real needs of users.

Therefore, an organized, strict and methodology-based assessment and evaluation of these solutions is then key in this field.

Considering all these aspects, the involvement of Living Labs in the evaluation process of smart living solutions can really make a difference, as it offers, among other advantages, the possibility to perform personalized evaluations, in real-life environments, in a particular and specialized ecosystem, and involving all the relevant stakeholders needed to convert an innovation into a real product or solution. Based on this, Living Labs can be considered really unique, and their characteristics allow them to guarantee the provision of the best personalized services in terms of the assessment of user needs, the evaluation of the usability aspects of a solution and the validation of a particular product in a relevant environment.

## 13.5 Examples of the Use of Living Labs in the Evaluation Process of Smart Living Solutions

### 13.5.1 *The Smart House Living Lab*

The Smart House Living Lab was created as part of the CIAMI project—Center for Experimental Research in Applications and Ambient Intelligence [18]. The project's mission was the development of Ambient Intelligence solutions covering all aspects related to prevention, care and promotion of health and well-being, social inclusion and independent life, and in general for all those people-oriented applications and services at all stages of the value chain: training, theoretical research, experimental research, technological development and technology transfer [6] (Fig. 13.1).

The most important and tangible result of the CIAMI project was the creation of the Smart House Living Lab, based in the Universidad Politécnica de Madrid and managed by the research group Life Supporting Technologies [22].

With an area of over 150 m<sup>2</sup>, this Living Lab features modern control technology, monitoring and regulation of the environment, where ideas are born, developed and the most cutting-edge services are tested in the e-inclusion (AAL) and e-health (p-Health).

The Smart House Living Lab consists of three distinct areas:

- The *User Area*, approximately 100 m<sup>2</sup> where we can find not only a digital home, kitchen, bathroom, bedroom and living room, but an open space where any scenario required can be simulated (e.g. a house, an operating room or an office). In these scenarios, the user can interact with the environment using different and diverse technologies embedded in the Living Lab (e.g. natural speech, touch screens, etc.) and also test and evaluate any additional solution integrated in this space.
- The *Control Room*, which has a unique view of the user through a one-way mirror. The room holds the communication systems with high bandwidth user area, the Internet and the server technologies. The environmental monitoring and interaction have been designed to be easily expandable and scalable in the future.



**Fig. 13.1** Smart House Living Lab facilities

- The *Area of Interaction in Virtual Reality*, used both for studying the user interaction with devices prior to prototyping and for training users in the use of new solutions.

In addition, the Smart House Living Lab provides an ecosystem of actors, such as a network of experts in several relevant fields (e.g. health technology assessment), national health system and universities. Within this ecosystem, the Living Lab performs the evaluation of products and services in a collaborative and efficient way, using a real environment infrastructure and accessing to end users. The feedback obtained from the end users contributes to achieve the goal of improving the products and services.

The combination of all these components makes the Smart House Living Lab an intelligent environment and ecosystem to evaluate new smart living solutions following co-design methodologies with end users.

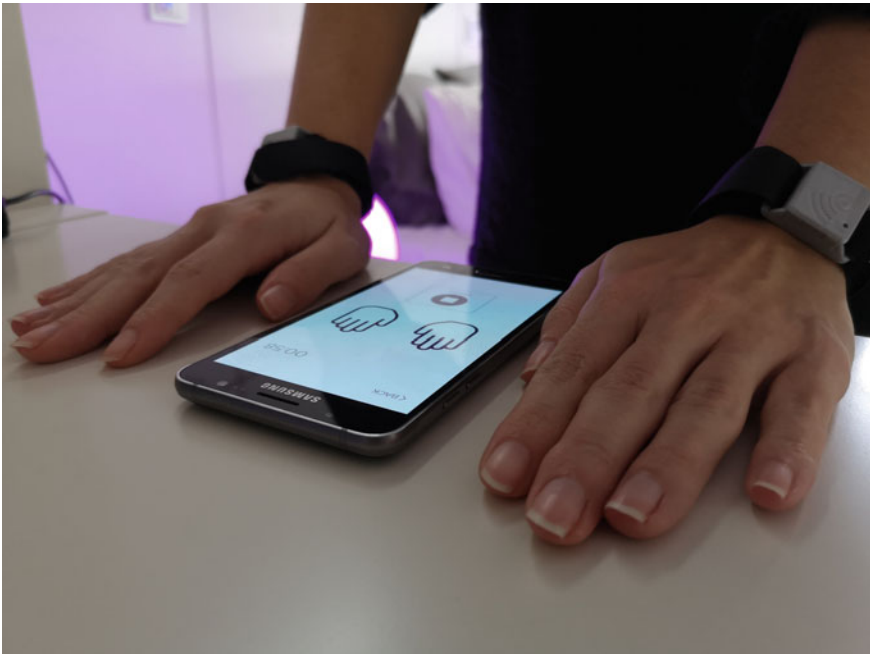
As an example of the work done in this Smart House Living Lab, we present here the collaboration in two different cases: HOOP, an EIT Health-funded project [13], and ACTIVAGE [1], a project partially funded by the European Commission under its Horizon 2020 framework programme.

### 13.5.2 HOOP Project Collaboration with Living Labs

HOOP was a project funded by the EIT Health that aimed to design and develop a tool for Parkinson's disease rehabilitation training at the patient's home, reducing the costs of usual therapies [13, 26].

The participation of the Smart House Living Lab in HOOP was essential, particularly in the design and execution of the evaluation study of the intelligent system developed within the project. The evaluation was planned, since the very beginning, following specific methods to ensure the acceptability and market acceptance of the final product. More specifically, the evaluation included the following steps:

1. Patient recruitment in Parkinson associations: patient engagement in the trial.
2. Test preparation with experts (psychologist, engineers and doctors)
3. Usability tests in Living Lab: testing with patients.
4. Feedback and coaching with the users (Fig. 13.2).



**Fig. 13.2** Usability test of HOOP systems with a user at Smart House Living Lab

### 13.5.3 *ACTIVAGE in the Smart House Living Lab*

ACTIVAGE is an H2020 project, partially funded by the European Union, that aims to design and implement solutions that, through the Internet of Things (IoT), enhance the independence, autonomy and well-being of older adults, by supporting and extending independent living in their daily environments. The project is implementing a large-scale multi-centred European pilot on intelligent living environments that will involve 10,000 people in nine Deployment Sites (DS) in seven European countries [1] (see Appendix for more information).

As one of its main results, the project plans to define an evaluation framework that includes, among others, the following indicators: quality of life, economic acceptability and usability. This should provide new tools for policymakers and other stakeholders for improving the decision-making process [1].

One of the project Deployment Sites, located in Madrid, has planned to use a user-centred design (UCD) approach. In this sense, the Smart House Living Lab is being a space of vital importance for the development of the project in its different stages: demonstration, expand and growth.

As it has been pointed out, the Smart House Living Lab is an accessible environment where new applications and technological services are co-created, tested and validated under the paradigm of ambient intelligence. In this sense, the Madrid DS of ACTIVAGE project has used these Living Lab facilities to perform two types of validation in the initial phase of the project:

- A technical evaluation, to test and assess the quality of the project's technological solutions.
- A first usability evaluation, to gather initial input and feedback from end users and caregivers.

Both evaluations were made with carers, end users and companies. In this way, it has been possible to identify four large groups of activities that were carried out in the Living Lab environment in the framework of the project:

- *Deployment activities.* This group of activities included the installation of the IoT universAAL platform [10] in the Smart House Living Lab servers, as well as the installation of the ACTIVAGE Center, and the rest of the devices and sensors, connecting them to the IoT universAAL platform and the technical tests of the integration of the ACTIVAGE Center. Currently, the technical tasks for the deployment and integration with AIoTES, the large IoT platform developed by the project, are being carried out in this environment.
- *Testing and validation.* Prior to each phase of deployment and operation with users, testing sessions were carried out on the different solutions that allowed the technical validation of the technologies and to obtain the input and feedback with the necessary adjustments to be made in order to reach an operational status in final versions.



Smart House Living Lab plays a relevant role in the innovation phases

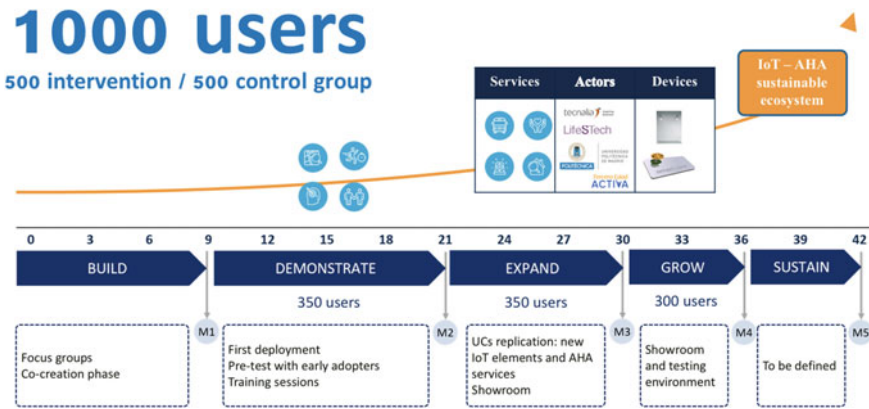


Fig. 13.3 Planning for ACTIVAGE MAD DS demonstration phase [1]

- *Support and installation procedures*, to ensure that the installers of the technologies had the necessary knowledge for the deployment, installation exercises and technical tests were carried out so that subsequent deployments involved as few errors as possible.
- *Training and coaching of users and caregivers*. Once the technical tests were completed, tests with user were carried out with 10 real end users. Also, caregivers were trained so that they could then introduce the technology to the end users.

Figure 13.3 depicts the planning for the Madrid DS demonstration phase, the first phase of the project.

In the next phases of the project—Expand and Grow—a similar approach will be followed to test the solutions in the Smart House Living Lab with the end user deployments.

### 13.6 Conclusions

In this chapter, we have presented the Living Labs as the new “instruments” to enhance the innovation process. Together with the actual concept of Living Lab, we have also described the current status of the different networks, their characteristics and some examples of the activities that a Living Lab can perform.

Living Labs are specialized environments that closely resemble real life. They use co-creation methods to understand the users need and, thus, help to develop better products and services. They also have access to end users and all the relevant stakeholders in a particular ecosystem, and they use different co-creation methods to produce evidence-based results. The combination of this set of characteristics into



a single organization makes the Living Labs unique and enables them to play an important role in the innovation field.

Although we have seen that the natural working space for Living Labs is the validation phase of the value chain, the truth is that they can also contribute significantly to earlier stages like the co-creation or ideation phase. Their deep knowledge of their local ecosystem and the needs of the particular type of end users, together with the use of UCD and co-creation methodologies make them perfect for helping other organizations along the innovation pathway.

In fact, one of the added values of the Living Labs lies in their potential role as a connector between the needs (the users) and the offer (the suppliers). In the end, all the co-creation and evaluation activities that they can do have the final goal of ensuring that the new solutions developed actually meet the needs of the end users and, thus, they have a potential market.

In summary, Living Labs are the perfect organizations to foster innovation and to contribute to launch new products and solutions to the market with the highest possibilities of success.

**Acknowledgements** The authors would like to thank the following projects for the valuable contributions to the elaboration of this chapter: EIT Health Living Labs and Test Beds project (2016–2019), EIT Health HOOP project (2017–2018) and Horizon 2020 ACTIVAGE project (H2020-732679). These projects have been partially funded by the EIT and the European Union.

## Appendix: The ACTIVAGE Project

ACTIVAGE is a large-scale multi-centred European pilot project on intelligent living environments that will involve 10,000 people in nine Deployment Sites (DS) in seven European countries. With a four-year duration, the project started in 2017 and is partially funded by the European Commission under its framework programme Horizon 2020 [1].

The objective of ACTIVAGE is to design and implement solutions that, through the Internet of Things (IoT), enhance the independence, autonomy and well-being of older adults, by supporting and extending independent living in their daily environments. This will promote the reduction of the impact caused by chronic diseases and age-related impairment and will also respond to the real needs of caregivers, service providers and policymakers [1].

ACTIVAGE aims to create evidence and to be a global reference engine to demonstrate that standard, safe and intra-operative IoT ecosystems can enable new business models and cost-effective solutions for active and healthy ageing. With this, it also aims at contributing to the sustainability of health and care systems, to the competitiveness of European industry through innovation, and to the improvement of the quality of life and autonomy of older adults in independent living. In this sense,

according to the project website “ACTIVAGE project will provide a set of techniques and tools as well as a co-creation framework that allows the identification, measurement, understanding and prediction of the demands and needs of the IoT ecosystem in Active Healthy Ageing (AHA) users: older adults, caregivers, professionals and health and social care providers, assessing their needs, preferences and perceptions regarding acceptance, trust, confidentiality, privacy, data protection and user safety” [1, 2].

ACTIVAGE project has worked from the beginning in analysing the socio-economic impact and sustainability of the systems, in order to validate new business, financial and organizational models for the provision of assistance [1].

## MADRID Deployment Site

Madrid (Spain) has an ageing population and a low rate of active ageing. Currently, in 2019, there are 650,000 older people living in the city and this number is expected to increase by 25% over the next 15 years. As a result, the number of people living in a situation of dependency is continuously increasing. For this reason, one of the nine Deployment Sites (DS) of ACTIVAGE project is located in this city and aims to reach 1000 users.

The current services related to active ageing in Madrid are offered through two types of institutions at the local and regional levels, as well as at the private and public levels, and they both present important gaps and limitations. This is the situation that the Madrid DS intends to transform through the implementation of the different solutions in the environment.

In total, three partners compose the Madrid DS: (1) Tercera Edad Activa SL, a company that provides services to elderly and disabled people; (2) Universidad Politécnica de Madrid, through its group Life Supporting Technologies, which are responsible of the development and integration of the IoT ecosystem and daily operations in the pilot; and (3) Fundación Tecnalia Research and Innovation, the providers of a technological service for balance assessment that is integrated as part of the solution offered to the used in the Madrid pilot. This group is deploying a set of innovative technological solutions—for the home and for the city—that allow working in four use cases linked to reality of the population aged 65+:

- Follow-up of people assisted outside the home.
- Promotion of exercise for fall prevention and physical activity.
- Cognitive stimulation for preventing mental decline.
- Prevention of social isolation.

## References

1. ACTIVAGE project (2017–2020) ACTivating InnoVative IoT smart living environments for AGEing well. Project partially funded by the European Union. H2020 Grant agreement No 732679
2. ACTIVAGE project website (2019) Retrieved from <http://www.activageproject.eu/>. Accessed 12 Apr 2019
3. Bajgier SM et al (1991) Introducing students to community operations research by using a city neighborhood as a living laboratory. Institute for Operations Research and the Management Sciences (INFORMS), Maryland, USA
4. Bergvall-Kareborn B, Hoist M, Stahlbrost A (2009) Concept design with a living lab approach. In: IEEE 42nd Hawaii international conference on system sciences. IEEE, Hawaii USA, pp 1–10
5. Chesbrough H et al (2008) Open innovation: researching a new paradigm. Oxford University Press, Oxford
6. Colomer J et al (2014) Experience in evaluating AAL solutions in living labs. *Sensors* 14(4):7277–7311
7. EIT Health (2019) EIT health website. <https://www.eithealth.eu/>. Accessed 12 Apr 2019
8. ENOLL (2019) European network of living labs. <https://enoll.org/>. Accessed 12 Apr 2019
9. ENoLL application guidelines (2019) ENoLL website. 13th wave application guidelines. <https://enoll.org/>. Accessed 12 Apr 2019
10. Ferro E, Girolami M, Salvi D et al (2015) The UniversAAL platform for AAL (Ambient Assisted Living). *J Intell Syst* 24(3):301–319. <https://doi.org/10.1517/jisys-2014-0127>
11. Forum LLSA (2019) Forum des Living Labs en Santé et Autonomie. <https://www.forumllsa.org/>. Accessed 12 Apr 2019
12. Halstad Living Lab (2019) Halstad Living Lab webpage. <http://www.halmstadlivinglab.se/>. Accessed 12 Apr 2019
13. HOOP project (2018) mHealth tOol for parkinson's disease training and rehabilitation at Patient's home. Project funded by the EIT Health, a programme supported by the European Union. Project number 18235
14. imec (2019) Living Lab Innovatrix © imec. <https://www.imec-int.com/en/business-model-innovation>. Accessed 12 Apr 2019
15. ISO/FDIS 9241-210 (2019) Ergonomics of human-system interaction—Part 210: human-centred design for interactive systems (under development in April 2019)
16. ISO 13407 (1999) ISO 13407 human-centred design processes for interactive systems
17. ISO 9241-210 (2010) ISO 9241-210:2010 Ergonomics of human-system interaction—Part 210: human-centred design for interactive systems
18. Montalva JB, Lazaro JP (2009) CIAMi: an experimental research centre for AmI applications and services. In: Proceedings of the DRT4All 2009, Barcelona, Spain, pp 18–25
19. Nesti G (2018) Co-production for innovation: the urban living lab experience. *Policy Soc* 37(3):310–325. <https://doi.org/10.1080/14494035.2017.1374692>
20. Nielsen J (1993) Usability engineering. Academic Press Inc, London
21. Picard R et al (2017) Co-design in living labs for healthcare and independent living: concepts, methods and tools. Wiley, Hoboken
22. Smart House Living Lab (2019) Smart house Living Lab (Life supporting technologies—Universidad Politécnica de Madrid). <https://www.lst.tfo.upm.es/smart-house/>. Accessed 12 Apr 2019
23. U.S Congress (1995) Innovation and commercialization of emerging technology. Office of technology assessment. OTA-BP-ITC-165, U.S. Government Printing Office, Washington, DC
24. Usability.gov (2019) <https://www.usability.gov/>. Accessed 12 Apr 2019
25. Van Geenhuizen M, Holbrook JA, Taheri M et al (2018) Cities and sustainable technology transitions. Leadership, innovation and adoption. Edward Elgar, UK
26. Villanueva-Mascato S et al (2018) A mobile system for PD patients based on music therapy. Published in IEEE EMBC 2018 proceedings

# Chapter 14

## A Distributed Spatial Index on Smart Medical System



Changqing Ji, Yang Gao, Zumin Wang and Jing Qin

**Abstract** Smart medical technologies, combine Internet of Things, cloud computing and artificial intelligence technologies, are redefining the family life. With the advent of the era of big data, traditional medical service systems cannot meet the needs of big data processing in the current medical system because of the limited computing resources, slow operation speed and poorly distributed processing capacity. In this chapter, cloud-based smart medical system applying MapReduce distributed processing technology is proposed to solve these problems. A new distributed  $k$ -nearest neighbour (kNN) algorithm that combines the Voronoi-inverted grid (VIG) index and the MapReduce programming framework is developed to improve the efficiency of the data processing. Here, VIG is a spatial index, which uses the grid structure and the inverted index based on Voronoi partitioning technology. The results of extensive experimental evaluations indicate the efficiency and scalability of the proposed approach with real and synthetic data sets.

**Keywords** Smart medical · MapReduce · Spatial · kNN

### 14.1 Introduction

With the development of the Internet of Things, big data and artificial intelligence, smart home makes life smarter. Smart cloud technology makes family life overcome time and distance constraints, and connects a single terminal device into a connected information network to extend or carry other value-added services. It not only solves the problem of home automation but also provides convenient services for family medical treatment, travel, diet and so on. In recent years, mobile medical technology has been an important technology in smart life.

---

C. Ji (✉)

Physical Science and Technology College, Dalian University, Dalian, China  
e-mail: [jcgood@gmail.com](mailto:jcgood@gmail.com)

Y. Gao · Z. Wang · J. Qin

Information and Engineering College, Dalian University, Dalian, China

© Springer Nature Switzerland AG 2020

F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications and Networks, [https://doi.org/10.1007/978-3-030-25590-9\\_14](https://doi.org/10.1007/978-3-030-25590-9_14)

Traditionally, people call the medical service hotline at home to obtain medical assistance. The patient cannot choose the medical institution to visit. With the continuous development of smart medical, mobile medical technologies such as electronic medical records, remote diagnosis and treatment, and wearable devices have been applied in the market [1]. It plays an important part in improving work efficiency, perfecting medical services and saving medical costs. More and more people are putting forward higher requirements for the quality and efficiency of medical services.

There is a big difference in medical service level between developed areas and remote areas because of the huge differences in geographical environment, uneven economic development and uneven distribution of medical resources [2]. We often come across emergencies that we do not know the location of the surrounding hospital, the hospital which can treat the pain with the better service, and the hospital which is closer. It will endanger the patient life without prompt treats and cures. In recent years, a large amount of spatial data is generated by GPS, Bluetooth and a large number of mobile applications based on location services and spatial query have emerged. We often encounter problems of long computing time and low efficiency when dealing with big data processing of mobile medical data. This is mainly due to the limited number of threads and storage capacity of traditional medical system. Introducing the concept of cloud and deploying medical applications to the cloud environment is a good solution to solve the above problems. The cloud-based smart medical system integrates multiple nodes to obtain distributed parallel computation ability and scalable storage capacity. It satisfies the needs of storing and managing large-scale medical spatial data and provides a fast response to query requests in the medical system.

This chapter will introduce a distributed spatial index and medical query algorithm based on smart medical system in cloud environment. The following is the contribution of the chapter.

- (1) We design the cloud-based smart medical system which includes the distributed spatial index and kNN query methods based on MapReduce framework supporting larger-scale medical spatial data.
- (2) We propose a combined index structure in cluster environment, which constructs grid index and Voronoi index, respectively. These two indexes are distributed indexes.
- (3) We implement spatial index based on MapReduce and optimise parallel kNN query algorithm based on this index structure. Experimental results show that our kNN query algorithm has better query performance.

This chapter extends the work in [3] and [4]. First, adding the motivation and background of the cloud-based medical system in detail. Secondly, further optimising the index structure of the past and introducing the grid index and how to implement the smart medical system. Finally, appending more experimental details to verify the scalability and performance of spatial index based on smart medical system.

## 14.2 Related Work

At present, many scholars use cloud computing technology to solve many problems in the field of smart medical. Zoc Doc [5] has launched an online booking platform based on location services for the US market. It seeks doctors based on geographic location, medical department, insurance and other conditions. Currently, it is used by about 700,000 people every month. Mc Gregor [6] proposed Artemis cloud services which provide real-time monitoring and clinical research for patients in remote areas. Lin et al. [7] proposed an integrated approach to quickly process, store, retrieve and analyse large-scale medical data. Nkosi et al. [8] proposed a cloud computing framework that enables mobile devices to be mitigated from overweight multimedia processing and security algorithms when delivering mobile healthcare services. Kayyali et al. [9] explained the impact of big data on the US health care field from multiple perspectives, pointed out that the combination of big data and the medical industry will generate a lot of potential value.

When considering scalable data processing, MapReduce [10, 11] has been the most popular technology. A data structure that can be constructed and processed in a distributed MapReduce environment is also a key consideration. There are several distributed index methods like R-tree [12, 13], Grid-based index [14, 15] and Voronoi [16]. However, the R-tree index which is a hierarchical index is not suitable to do parallelisation for MapReduce. The grid-based index structure in a distributed environment can cause to redundancy of content.

Existing algorithms that are used to solve complex query include kNN [17], Skyline [18, 19] and RkNN. The kNN algorithm is the most important method in these query algorithms. The existing principal problem is how to design distributed index in smart medical system to optimise kNN queries of large-scale spatial data.

The inverted index (inverted files) [20] is extensively used in the field of similarity searches, machine vision and retrieval. Zhang et al. [21] proposed H-BRJ and approximate H-zkNNJ algorithms in MapReduce to achieve effective parallel kNN join on the amount of data which is better than baseline methods. Pan [22] proposed GPU-based LSH algorithm in parallel environment, and this method achieved common kNN algorithm in high-dimensional area. Subsequently, Aleksandra et al. [23] introduce a method which implements the LSH-based MapReduce, and it also plays the data in extremely high-dimensional area. There are also many studies that are similar to our methods, where kNN queries were implemented by MapReduce, by using the inverted-LSH [24] and the inverted grid [14] methods, respectively. Akdogan et al. partition with the Voronoi diagram in MapReduce framework to answer range queries and kNN queries [16]. However, there are still many drawbacks. They need much space expense, and the performance of query is not good.

We mainly focus on data structure which can solve the problem of spatial kNN queries in cloud-based smart medical system.

## 14.3 Preliminaries

In this section, we first introduce the framework of the smart medical system. Then introduce the grid index and MapReduce framework. Finally, define kNN query and Voronoi Diagram.

### 14.3.1 Smart Medical System

The smart medical system consists of six parts, real-time user request module, smart medical application interface module, inverted spatial index maintenance module, query module, scheduling module and resource feedback module.

Among them, the resources feedback module is referred to service vehicle or the service people who can receive the task. The amount of data information generated by the system module will usually be relatively large. Besides, both its state and the medical spatial data will be continuously updated. As shown by the dotted arrows, the resources feedback system maintains spatial index of the resource for the purpose of fast user query processing. At the same time, it will continually update its own multidimensional data information (including temporal and spatial data). In addition, the service centre is responsible for continuous monitoring and updating the information of service request made by users, the information of whether the assigned resources have complete the service or not, and other related information status. After connecting the system, these two types of information is regularly communicated to the service centre through the medical system application interface within the time window (for example, once every 40 min).

As shown in Fig. 14.1, a user submits a medical service query through the smart medical system application interface to the system and receives a response from the service. As demonstrated by the solid arrows, all incoming queries of the system are streamed into a distributed cache queue and are processed according to the first-come-first-serve principle.

For each at the top of the query result queue return from distributed data processing system, the system invokes the searching module to search for a set of candidate result set which is likely to satisfy the query based on the latest inverted index.

### 14.3.2 Grid Index

In order to implement kNN query of spatial data in smart medical system, we have carried out a lot of research and proposed VIG index. We build a two-layer index in cluster environment. The index consists of two parts, the Voronoi index and the

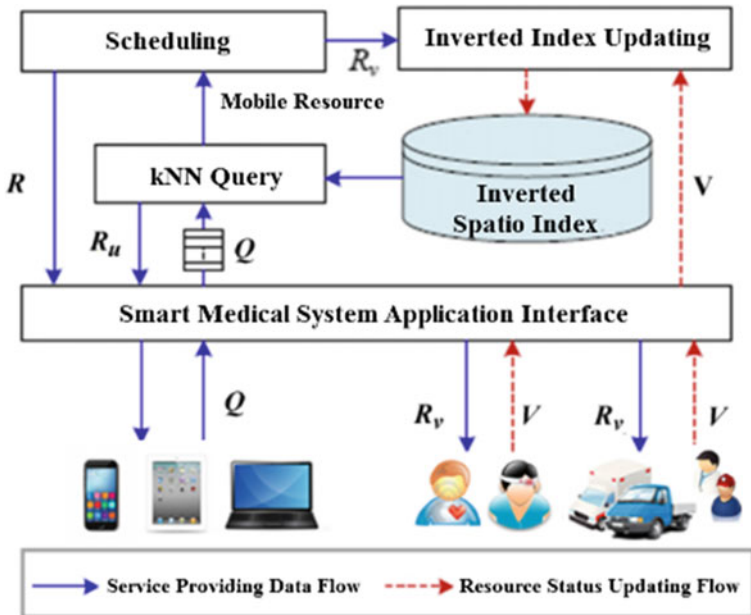


Fig. 14.1 Framework of mobile medical call system

grid-index-based inverted form. We partition spatial data based grid and build a grid-based index at the bottom of the node. Each data object is divided into a fixed-size grid. In order to improve the efficiency of the kNN query, we must ensure that the spatial data with similar physical locations are similar in the cluster. Therefore, we partition and store data using Voronoi partitioning technology. As the result, the data in the Voronoi partition of the query point  $P$  and the Voronoi partition adjacent to it must be the nearest neighbour of the query point  $P$ . However, the boundary treatment of the Voronoi partition for excessive amount of data may be wasted too much time. Therefore, we elect the clustering centre of the dataset using pre-clustering technology before the partition stage. This method reduces the number of spatial data objects greatly and saves a lot of time to build partitions.

The idea of grid indexing is to divide the data area into several smaller grids according to certain rules, each of which contains smaller grids or data ids. A grid index generally creates one or more index files, and the value of each index is the spatial data ID. In fact, the method avoid traversing in the whole data set can speed up the query efficiency.



### 14.3.3 *k*NN

The data objects were considered in an  $n$ -dimensional space  $D$ . Given two data objects  $p_1$  and  $p_2$ ,  $\text{Dist}(p_1, p_2)$  represents the distance between  $p_1$  and  $p_2$  in  $D$  as (14.1). The Euclidean distance is used as distance measure in the rest of this chapter, i.e.

$$\text{Dist}(p_1, p_2) = \sqrt{\sum_{i=1}^n (p_1[i] - p_2[i])^2} \quad (14.1)$$

where  $p_1[i]$  (resp.  $p_2[i]$ ) denotes the value of  $p_1$  (resp.  $p_2$ ) along the  $i$ th dimension in  $D$ . If  $p_1$  and  $p_2$  are not objects, the distance definitions may be needed to be defined by the user (Such as metric distance).

A  $k$ NN query is formally defined as below:

**Definition 1**  $k$ NN

Given a point  $q$ , a dataset  $S$  in space  $D$  and an integer  $k$ , the  $k$ -nearest neighbours of  $q$  form  $S$ , denoted as  $k\text{NN}(q, S)$ , is a set of  $k$  point from  $S$  that,

$$\forall p \in k\text{NN}(q, s), \forall s \in S - k\text{NN}(q, s), \text{dist}(p, q) \leq \text{dist}(s, q) \quad (14.2)$$

**Definition 2**  $k$ NN queries

Given two dataset  $R$  and  $S$  in space  $D$ , and an integer  $k$ .  $k$ NN queries of  $R$  and  $S$  [denoted as  $k\text{NN}(R, S)$ ], combine each point  $q \in R$  with its  $k$ -nearest neighbours from  $S$ .

$$k\text{NN}(R, S) = \{(q, k\text{NN}(q, s)) | q \in R\} \quad (14.3)$$

### 14.3.4 *MapReduce Framework*

Hadoop is a software platform for the development and operation of large-scale data. It is an open-source framework implemented by Java language, which realises distributed computing of large-scale data in a cluster [25]. HDFS and MapReduce are the core of the Hadoop framework.

Hadoop creates a task for each split. The map will output the result as a key-value pair. Hadoop is in charge of mapping the map by key-value pairs. The output is sorted and used as the input of Reduce. The output of the key-value pair is saved on HDFS as the final result.

The Hadoop cluster is mainly composed of NameNode, DataNode, Secondary NameNode, JobTracker and TaskTracker. The NameNode records how to split the data set into several blocks. The blocks are stored in the DataNode. The Secondary NameNode collects status information about the operation of file system. As a manager, JobTracker divides a job into multiple sub tasks, and then manages the schedule to be executed by TaskTrackers.

### 14.3.5 Voronoi Diagram

In this part, we will introduce the Voronoi diagrams.

In the geographic analysis, the Voronoi diagram [26] is often used for the interpolation of geographic entities and the analysis of influence area. It is another common tool to solve the problem of the nearest neighbour. The Voronoi is used to divide the spatial data into multiple polygon, named Voronoi cell. A data point  $p$  is closest to any point in common area, and there is no need to calculate the distance.

#### Definition 3 Voronoi Cell (VC)

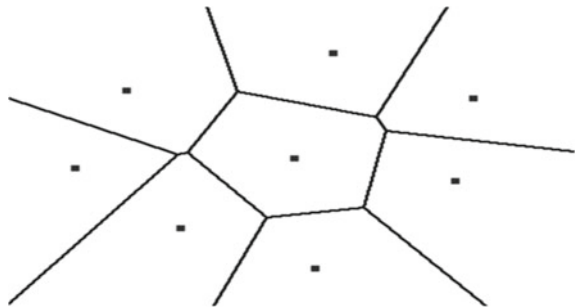
Consider  $P\{p_1, p_2, \dots, p_n\}$  of  $n$  points in space  $R$  partitions, the space of  $D$  into a regions, where  $n \geq 2$  and  $p_i \neq p_j$  for  $i \neq j, i, j \in I_n = 1, \dots, n$ . The region given by  $VC(p_i) = \{p | d(p, p_i) \leq (p, p_j)\}$  where  $d(p, p_i)$  express the minimum distance by Euclidean between  $p$  and  $p_i$ . It is named as Voronoi Cell (VC). Figure 14.2 depicts the Voronoi diagram of eight points in two-dimensional Euclidean space.

In Sect. 14.4, we will introduce the detail of our spatial index and kNN query algorithm.

## 14.4 Handling Spatial Index and KNN Using MapReduce

Our research contains a few assumptions. First, the data is too much to imagine so that it is inefficient for an internal memory calculating model to process. Second, we

Fig. 14.2 Voronoi diagram



adjust the data set to make sure the query point not found in dataset which is random occurrence. Thirdly, the data model is established on the basis of the multi-dimension Euclidean space and distance.

In this section, firstly, giving a brief overview of the inverted Voronoi index. Then introducing the grid-based index. Thirdly, describing how to build our VIG index in MapReduce framework. Finally, introducing how to optimise the kNN algorithm based on VIG index.

### 14.4.1 Inverted Voronoi Index Structure

In full-text search, the most commonly used is the inverted index structure [27]. The general index structure is to find the location of the record, and then match the keywords what we looking for. Inverted index is just the opposite. It determines the location of the record through the keywords that we look for, rather than determining the required keywords through the record. In this chapter, the idea is to map the partition number of Voronoi diagram and object id.

**Definition 4** Inverted Voronoi Index (IVI)

Providing a large-scale dataset  $p$ , which includes a lot of data objects in Euclidean space. Each object is divided into a Voronoi Cell (VC). Voronoi diagram can be recorded by  $VC(p) = \{VC_1, VC_2, \dots, VC_m\}$ . We make the  $VC(p)$  as the key of the inverted index. All ids of the data objects which satisfy  $\{P_i\} \in VC_m$  are sorted into lists as the value of the inverted index. That is, each VC contains a list of object ids.

Figure 14.3 constructs the IVI for 2D spatial data objects in six polygon by the partitioning algorithm of Voronoi diagram. In a word, make  $p$  become the leader of the set of points.  $\forall p_i \in p$ . Thus,  $P_i$  is appointed to the VC polygon with its closest leader  $P$ .

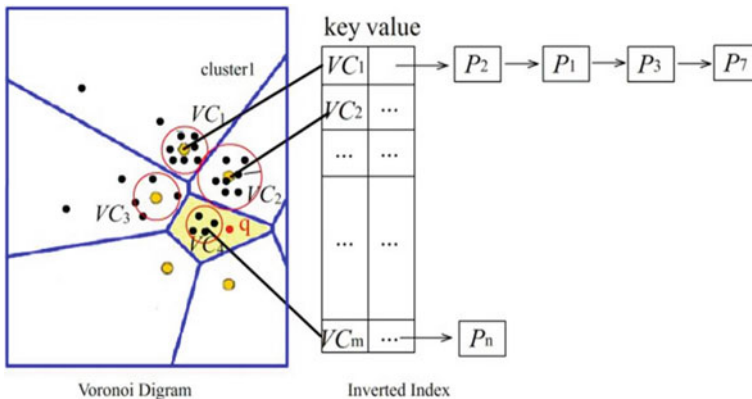
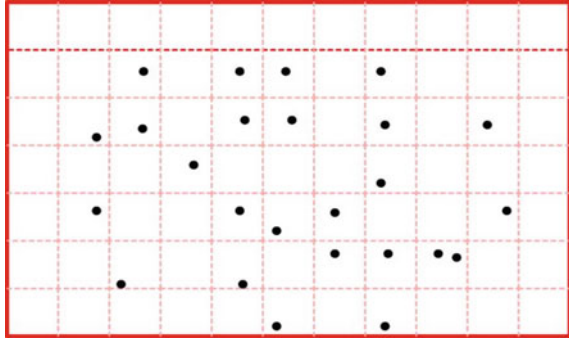


Fig. 14.3 Inverted Voronoi index

**Fig. 14.4** Grid-based index

### 14.4.2 Grid-Based Indexing

The underlying index of our smart medical system uses a grid index. Grid index is commonly used in the spatial data index. As shown in Fig. 14.4, the plane is split into uniform grids according to a certain width and height. Within each grid cell, we choose the points with special significance, such as road network node, which is closest to the geographical centre of the cell as the anchor node of the cell (represented by a black dot in Fig. 14.4).

For each grid element, as a collection object, record information (address or reference) of all the primitives contained in this grid. In this chapter, the query condition is composed of key, and the value is mapped to the location of the local data node.

### 14.4.3 Voronoi-Inverted Grid Index

In this chapter, we use Voronoi to partition the grid data and construct the inverted index structure based on the division of the distribution of data according to the actual distribution in the two-dimensional quadrant. The system based on key-value form, with the Voronoi cell of cluster centre point as the key and the data node area cell (DC) of the surrounding data point as the value to establish the inverted grid index structure as Fig. 14.5.

#### **Definition 5** Voronoi-Inverted Grid Index (VIGI)

Voronoi-inverted grid index is a large-scale grid data structure that storage data points. Given a large-scale dataset  $P$  which contains a series of points in Euclidean space. For indexing a data set, each object is divided into a Data Node Areas Cell,

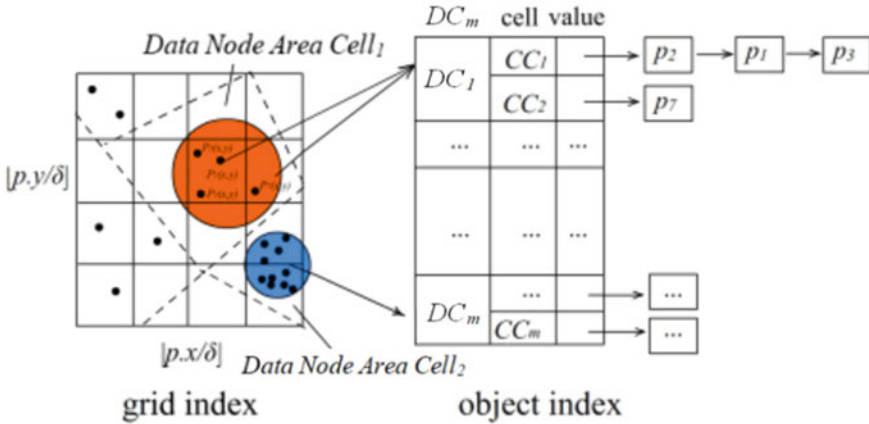


Fig. 14.5 Inverted spatial index

Data Node Area Cells can be recorded by  $DC(p) = \{DC_1, DC_2, \dots, DC_m\}$ . We make the  $DC(p)$  as the key of the VIG index. All IDs of the data objects which satisfy  $\{P_i\} \in DC_m$  are sorted into lists as the value of the VIG index. That is, each DC contains a list of objects.

### 14.4.4 Constructing VIGI with MapReduce

Inverted index [20] is proved to be an effective structure. In this section, we describe the method how to build the VIG with MapReduce. The construction of VIG is suitable for parallel computing, because it can be obtained by merging multiple Voronoi partitions.

Algorithm 1 shows how to construct the spatial index with MapReduce. First, we take two dataset  $R$  and  $S$  in the  $d$ -dimensional space as input splits by the Hadoop default mechanism. It mainly includes Map tasks and Reduce tasks, which is generated by us. Before performing the map task, we use the fast pre-clustering algorithm to get some cluster centres as the representative point  $P$  of each Voronoi partition and load into the main memory of each mapper. According to the pre-clustering results, we have manually divided the boundaries of the Voronoi map, marked each cell as  $\{DC_1 \dots DC_m\}$ .

**Algorithm 1:** Constructing VIGI using MapReduce**Input:** Dataset  $S$ **Output:** Voronoi Inverted Grid Index**Map (k, v)** //Map task

1. **for** each point  $r \in R, s \in S$  and pivots  $P$  in  $R$  and  $S$  in the dataset;  $o$  is not null; **do**
2. Calculate the  $\text{Min}(\text{dist}(o, P))$
3. Assigns  $o$  to the closest pivot  $P \in DC_m$ ;
4. **if**  $DC_m$  is not null **then**
5. Assigns  $o$  to the grid  $o \in CC_m$
6. emit  $\langle\langle DC_m, CC_m \rangle, \text{List}(P_i) \rangle$  of  $R$  and  $S$ ;
7. **end if**
8. **end for**

**Reduce (k, v)** //Reduce task

9. sort the objects in the  $\text{List}(p_i)$  of  $R$  and  $S$ ;
10. emit the List of  $(P_n)$ ;
11. Output key-value pairs  $\langle DC_m, CC_m \rangle, P_n$ ;

Then, in each mapper, it reads the input splits (which is always associated with blocks in the distributed file system) sequentially through the `TextInputFormat`. The `TextInputFormat` defines how the data in the file is read into the Mapper instance. It calculates the distance between all points of object  $r, s$  and all pivots in  $P$ , and assigns  $r, s$  to the nearest pivot  $P$ . Each point is gathered together in the Cells in lines 2–3,  $m$  Cells will be generated, and then separately emit the  $\langle\langle DC_m, CC_m \rangle, \text{List}(P_i) \rangle$  pairs in lines 4–8. The mapper outputs each object  $r, s$  along with its partition id  $DC_m$ , the original dataset names ( $R$  or  $S$ ), and the distance to the nearest pivot. Finally, we need to write the output of the mapper to the HDFS using our own `MultipleOutputFormat` to meet our own requirements, which determines how to write the results of a job back to the underlying persistent storage in lines 9–11.

#### 14.4.5 The $k$ NN Algorithm Based on VIGI Using MapReduce

In this section, we introduce a distributed query operator by applying the concept of Voronoi-inverted grid index based on MapReduce framework.

Given a query object  $q$ , we execute the  $k$ NN using the index our proposed named VIGkNN. The VIGkNN algorithm consists of two phases; the first phase is collecting candidate objects. Given the query point  $q$ , we query the pivot objects with VIGI that are designed. And then, transfer the candidate objects as the input of the second phase. The second phase determines the query results among the candidate objects by MapReduce distributed parallel framework.

In VIGI, we have to build the projection of point sets. We also have determined the partitioning value using Voronoi diagram for point sets. According to the definition of  $k$ NN query in Sect. 14.3, two datasets  $R$  and  $S$ , and an integer  $k$  is given.  $k$ NN( $R$ ,

$S$ ) combine each point  $q \in R$  with its  $k$ -nearest neighbours from  $S$ . Therefore, we need to partition the  $R$  and  $S$  based on Voronoi-inverted grid index in order to parallel kNN queries using MapReduce.

---

**Algorithm 2:** VIGkNN
 

---

**Input:** Dataset  $R, S$

**Output:** VIGkNN( $R, S$ )

**Map** ( $k, v$ ) //Map task

1. **if** dataset  $R_i \in R$  **then**
2.  $DC_m \leftarrow \text{getPartitionID}(R.\text{partition});$
3.  $CC_m \leftarrow \text{getgridID};$
4.  $\text{output}(\langle DC_m, CC_m \rangle, (k, v));$
5. **else**
6. **for** each pivots  $P$  in  $S$  **do**
7. **if**  $\text{dist}(s, P) < \text{dist}(r, P)$  **then**
8.  $\text{pid} \leftarrow \text{getPartitionID}(S.\text{partition});$
9.  $\text{output}(\langle DC_m, CC_m \rangle, (k, v));$
10. **end if**
11. **end for**
12. **end if**

---

**Reduce** ( $k, v$ ) //Reduce task

13. parse  $P$  from  $\langle \langle DC_m, CC_m \rangle, (k, v) \rangle$
  14. compute the kNN( $r, S$ )
  15. Output key-value pairs  $\langle r, \text{kNN}(r, S) \rangle$
- 

As shown in Algorithm 2, we first place the file containing partition values for  $R$  and  $S$  based on VIGI. Then, the master loads the file to the distributed cache. Mappers read partition values for each split of  $R_i \in R$  and  $S_i \in S$  from the distributed cache to generate key-value pairs. The map function generates a new key-value pair for each object  $r \in R$ , the key of which is the partition id as well as the value forms  $k$  and  $v$  (lines 1–4). The map function also forms a series of new key-value pairs for each object  $s \in S$ , if  $\text{dist}(s, P) < \text{dist}(r, P)$  (lines 6–11). By doing this, objects in each partition of  $R$  and the possible  $k$ -nearest neighbours will be sent to the similar reducer. Then, mapper writes to distributed file system. Each reducer iteratively reads all points within a pair of  $R_i$  and  $S_i$  received from the mappers, presents the certain task by a group of pairs with the same key of  $DC_m$  and performs the kNN query. Finally, after checking all partitions of  $S_i$ , the reducer outputs kNN( $r, S$ ) and key-value pairs  $\langle r, \text{kNN}(r, S) \rangle$  (lines 14–15). The algorithm outputs the key-value pairs  $\langle r, \text{kNN}(r, S) \rangle$  to get the results of kNN queries.

**Table 14.1** Description of the spatial data sets

| Data Set             | Objects (number) | Description                                    |
|----------------------|------------------|--|
| RDS (Real data)      | 1,500,000        | Ambulance GPS data with medical system of city |
| SDS (Synthetic data) | 4,700,000        | Follow uniform and Zipf distributions          |

## 14.5 Experimental Study

In the following part, we will examine the performance of the index and the algorithm. We first introduce the experimental data set, the default parameters and the settings of the experimental environment. Next, we give the analysis of the experimental results.

### 14.5.1 Experiment Setting

We build a cluster by a series of 32 server PC machines. Every node holds a Dual Core AMD 2.00 GHz CPU, 8 GB memory and 73 GB of SCSI hard drive. The operating system is Ubuntu 10.10. Hadoop version is Hadoop0.20.2. Each slave nodes have one TaskTracker and DataNode. The master node has a single NameNode and JobTracker.

In order to better adapt the Hadoop environment, we make some changes of the default Hadoop configurations: (1) the replication factor is set to 1; (2) each node is set to run one map and one reduce task; (3) the size of virtual memory for each map and reduce task is set to 4 GB.

We take the ambulance GPS data with medical system as the real-world data set (RDS) in the experiment [28], including seven-dimensional for approximately 1,800,000 data points. The original data set is about ten gigabytes before being decompressed. In this experiment, we selected a subset of 1,500,000 data points, in which every data point represents the parking behaviour of the vehicle.

At the same time, we use a random generator to generate a series of data as the experimental simulation data set (SDS), which follows the uniform distribution and Zipf distribution. The dimension changes from 2 to 5, and the base number is from 128 to 4096 k (i.e. more than 4 million points). The data set is described in Table 14.1.

### 14.5.2 Performance Evaluation

In this part, we compared Voronoi-inverted grid index based on MapReduce (MRVIG), Inverted Voronoi index based on MapReduce (MRIV), R-tree index based on MapReduce (MRTree) [13] and Voronoi index based on MapReduce (MRV) [16].



In this part of the experiment, we first compare the efficiency of constructing the MRVIG, MRIV, MRTree and MRV index. This experiment compares the creation time of the four indexes by changing the number of nodes from 2 to 32. Figure 14.6 depicts the results of the node number when constructing index on RSD and SDS datasets.

As the number of cluster nodes grows, the creation time of the four index structures decreases almost linearly in Fig. 14.6. In addition, the construction time of MRVIG is the shortest compared with the MRTree, MRV and MRIV in the similar node number. While the node numbers is 8 and 16, MRVIG is 50% faster than MRIV, six times faster than MRTree, and at least four times faster than MRV. Also, we can know that MRIV is at least three times faster than MRV and four times faster than MRTree. Because the structure of Voronoi-inverted grid index and Voronoi index is simple. It takes a lot of time to build an index of the R-tree structure. This hierarchical tree structure is much more complicated than Voronoi index. Due to the methods of pre-clustering, we only need to construct the Voronoi diagram for pivots. Therefore, the construction of MRIV and MRVIG index needs less time.

In this experiment, we contrast the query time and scalability of VIGkNN, IVkNN and Voronoi-based method (VkNN). Figure 14.7 shows the outcomes of different data sizes  $n$  of RDS and SDS, the response time of kNN query is not notably increased. The VIGkNN methods have a better performance than VkNN and IVkNN when adding more physical slaves to the cluster.

In the last part of experiments, we examine the effect of different  $k$  on the performance of technology we proposed. Figure 14.8a, b show the results of comparison between VIGkNN, IVkNN and VkNN methods, respectively, by changing  $k$  from 5 to 100 on both the RDS and SDS. As Fig. 14.8 shows, our VIGkNN method takes less time, and the query performance is slightly better than IVkNN and significantly better than Voronoi-based indexing. For small  $k$  values, a determinant of performance that must be the speed of kNN query; while for large  $k$  values, the communication overhead will gradually become the determinant in the IVkNN and

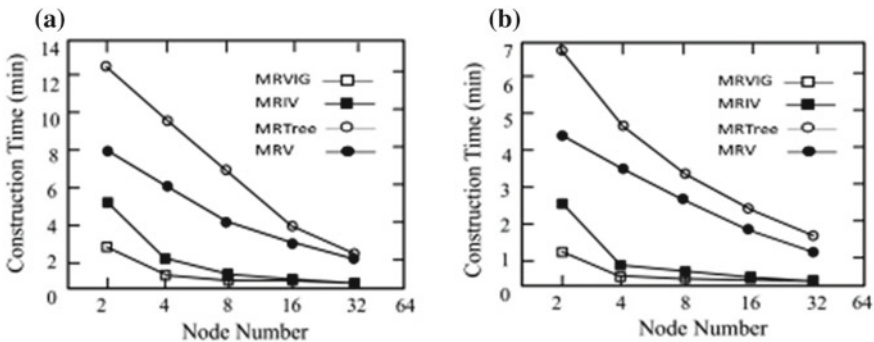
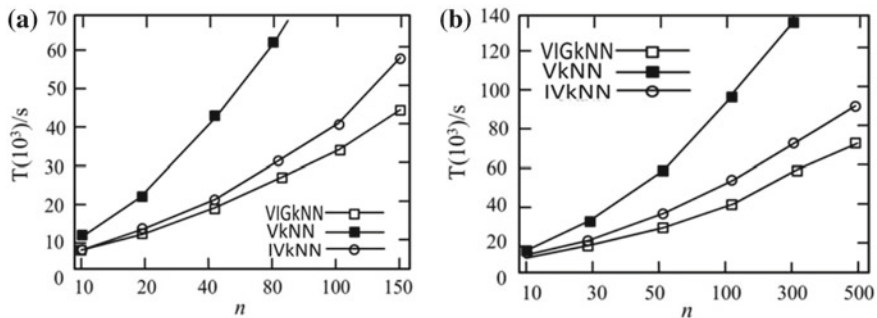
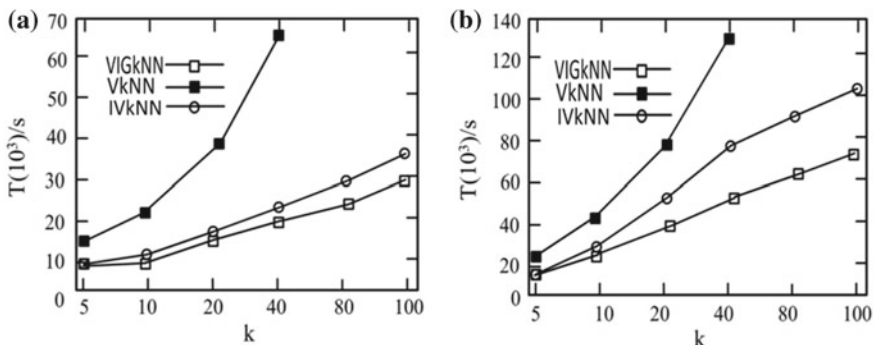


Fig. 14.6 Construction time of MRVIG and MRIV and MRTree and MRV on real data set and synthetic data set. a Real data set. b Synthetic data set



**Fig. 14.7** Execution time of VIGkNN, VkNN and IVkNN method efficiency of kNN on real data set and synthetic data set. **a** Real data set. **b** Synthetic data set



**Fig. 14.8** Execution time of VIGkNN, VkNN and IVkNN with different  $k$  on real data set and synthetic data set. **a** Real data set. **b** Synthetic data set

VkNN methods. The experimental results show that the communication overhead can be effectively reduced through our method.

## 14.6 Conclusions

This chapter proposed a smart medical system supporting spatial kNN query. We proposed a new distributed spatial index: the VIG index, which combines the inverted index and the spatial partition technology. Adding a design based on pre-clustering and Voronoi, our implementation layers use grid and inverted index structures. The experiment demonstrates that it can support the medical query well in distributed cloud environment.

The smart medical system based on Hadoop, a popular open-source distributed cloud computing framework. The scalability and efficiency of the design we proposed

are proved through a experimental evaluation. The experiment proves the scalability of VIG by adjusting the number of cluster nodes while serving multi-dimensional data sets kNN query in real time with execution times. In the future, we will extend our research by adding more efficient and convenient index method such as secondary index and double-level index to solve the shortboard of large-scale high-dimensional complex operators such as skyline or reverse NN.

**Acknowledgements** This work is supported by the National Nature Science Foundation of China (61702071 and 61501076), Industrial and educational cooperation and collaborative education project of the ministry of education (201702029010), Science and Technology Public Welfare Institute Fund of Liaoning Province (20170053), the Project of College Students' Innovative and Entrepreneurial Training Program (2018004, 2018016, 2018044, 2018046, 2018066, 2018125, 2018173, 2018243), the Key Research and Development Program of Liao Ning Province of China (2017104014), the Liao Ning Provincial Ph.D. Start-up Foundation of China (20170520438), the Natural Science Foundation of Liao Ning Province of China (20180551247), the Science and Technology Innovation Fund Project of Dalian of China (2018J12GX049, 2018J13SN088), Dalian Key Laboratory of Smart Medical and Health, Liaoning optoelectronic information technology engineering laboratory.

## References

1. Li J (2015) Design and implementation of a mobile- health nursing call system based on cloud computing. Xian, China
2. Ding H, Moodley Y (2012) A mobile-health system to manage Chronic Obstructive Pulmonary Disease patients at home. In: 34th annual international conference of the IEEE EMBS, San Diego, CA, 28 Aug–1 Sept 2012
3. Ji C, Wang B, Tao S et al (2016) Inverted Voronoi-based kNN query processing with MapReduce. In: Trustcom/BigDataSE/I SPA, IEEE, Tianjin, China, 23–26 Aug 2016
4. Gao Y, Wang Z, Ji C et al (2017) Design and implementation of a mobile-health call system based on scalable kNN query. In: e-Health networking, applications and services (Healthcom), 2017 IEEE 19th international conference on, IEEE, Dalian, China, 12–15 Oct 2017
5. Winter A, Stäubert S, Ammon D et al (2018) Smart medical information technology for health-care (SMITH). *Methods Inf Med* 57(01):e92–e105
6. Mc Gregor C (2011) A cloud computing framework for real-time rural and remote service of critical care. In: IEEE symposium on Computer-Based Medical Systems, Bristol, 27–30 June 2011
7. Lin C, Huang L, Chou S et al (2014) Temporal event tracing on big healthcare data analytics. In: Proceedings of 2014 IEEE international congress on Big Data. IEEE, Anchorage, 27 June–2 July 2014, pp. 281–287
8. Nkosi M, Mekuria F (2010) Cloud computing for enhanced mobile health applications. In: 2010 IEEE second international conference on Cloud Computing Technology and Science (CloudCom), Indianapolis, 30 Nov–3 Dec 2010
9. Kayyali B, Knott D, Van Kuiken S (2013) The big-data revolution in US health care. Accelerating value and innovation. *Mc Kinsey* 2(8):1–13
10. Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters. *Commun ACM* 51(1):107–113
11. The Apache Hadoop Project (2010) <http://hadoop.apache.org/core/>
12. Hesabi Z, Sellis T, Liao K (2018) DistClusTree: a framework for distributed stream clustering. In: Australasian Database Conference. [https://doi.org/10.1007/978-3-319-92013-9\\_23](https://doi.org/10.1007/978-3-319-92013-9_23)

13. Cary A, Sun Z, Hristidis V et al (2009) Experiences on processing spatial data with MapReduce. In: International conference on Scientific and Statistical Database Management. Springer. [https://doi.org/10.1007/978-3-642-02279-1\\_24](https://doi.org/10.1007/978-3-642-02279-1_24)
14. Ji C, Dong T, Li Y et al (2012) Inverted grid-based kNN query processing with MapReduce. In: 2012 Seventh ChinaGrid annual conference, IEEE, Beijing, 20–23 Sept 2012
15. Chen L, Cox S, Goble C et al (2002) Engineering knowledge for engineering grid applications. In: Proceedings of Euroweb 2002 conference
16. Akdogan A, Demiryurek U, Banaei-Kashani F et al (2010) Voronoi-based geospatial query processing with MapReduce. In: Cloud computing technology and science (CloudCom), IEEE, Indianapolis, 30 Nov–3 Dec 2010
17. Gonzalez-Lopez J, Ventura S, Cano A (2018) Distributed nearest neighbor classification for large-scale multi-label data on spark. *Future Gener Comput Syst* 87:66–82
18. Li Y, Li Z, Dong M et al (2015) Efficient subspace skyline query based on user preference using MapReduce. *Ad Hoc Netw* 35:105–115
19. Choi H, Lee M, Lee K (2012) Distributed high dimensional indexing for k-NN search. *J Supercomputing* 62(3):1362–1384
20. Zobel J, Moffat A, Ramamohanarao K (1998) Inverted files versus signature files for text indexing. *ACM Trans Database Syst* 23(4):453–490
21. Lu W, Shen Y, Chen S et al (2012) Efficient processing of k nearest neighbor joins using MapReduce. *Proc VLDB Endowment* 5(10):1016–1027
22. Pan J, Manocha D (2011) Fast GPU-based locality sensitive hashing for k-nearest neighbor computation. In: Proceedings of the 19th ACM SIGSPATIAL international conference on Advances in Geographic Information Systems, Chicago, Illinois, 1–4 Nov 2011
23. Stupar A, Michel S, Schenkel R (2010) RankReduce—processing k-nearest neighbor queries on top of MapReduce. In: Workshop on Large-Scale Distributed Systems for Information Retrieval, Geneva, Switzerland
24. Zhu P, Zhan X, Qiu W (2015) Efficient k-nearest neighbors search in high dimensions using MapReduce. In: 2015 IEEE fifth international conference on Big Data and Cloud Computing (BDCloud), IEEE, Dalian, 26–28 Aug 2015
25. Welcome to Apache Hadoop!. [hadoop.apache.org](http://hadoop.apache.org). Accessed 16 Dec 2015
26. Sack J, Jorge U (eds) (1999) Handbook of computational geometry. Elsevier
27. Comap, the consortium for mathematics and its applications, <http://www.comap.com>. Accessed 26 Oct 2011
28. Lian X, Chen L (2008) Probabilistic group nearest neighbor queries in uncertain databases. *TKDE* 20(6):809–824

# Chapter 15

## Data Reduction Methods for Life-Logged Datasets



William P. Burns, Paul J. McCullagh, Dewar D. Finlay,  
Cesar Navarro-Paredes and James McLaughlin

**Abstract** Life-logging utilises sensor technology to automate the capture of a person's interaction with their environment. This produces useful information to assess wellbeing, but this information is often buried within the volume of data. In this chapter, we analyse a life-log comprising image data and contextual information and apply algorithms to collate, mine and categorise the data. Four approaches were investigated: (i) Self-reporting of important events by the person who collected the data; (ii) Clustering of images into location-based events using GPS metadata, (iii) Face detection within the images and (iv) Physiological monitoring using Galvanic Skin Response (GSR); as a way to identify more meaningful images. Using a bespoke wearable system, comprising a smartphone and smartwatch, six healthy participants recorded a life-log in the form of images of their surroundings coupled with metadata in the form of timestamps, GPS locations, accelerometer data and known social interactions. Following approximately 2.5 h of recording, the data reduction methodologies outlined above were applied to each participant's dataset, yielding an 80–86% reduction in size which facilitates more realistic self-quantification. However, each approach has some shortcomings and the data reduction method used will need personalisation and depend on the intended application.

**Keywords** Life-log · Data reduction · Self-report · Geo-data mining · Face detection · Physiological · GSR

---

W. P. Burns (✉) · D. D. Finlay · C. Navarro-Paredes · J. McLaughlin  
School of Engineering, Ulster University, Newtownabbey, Northern Ireland, UK  
e-mail: [wp.burns@ulster.ac.uk](mailto:wp.burns@ulster.ac.uk)

D. D. Finlay  
e-mail: [d.finlay@ulster.ac.uk](mailto:d.finlay@ulster.ac.uk)

C. Navarro-Paredes  
e-mail: [c.navarro@ulster.ac.uk](mailto:c.navarro@ulster.ac.uk)

J. McLaughlin  
e-mail: [jad.mclaughlin@ulster.ac.uk](mailto:jad.mclaughlin@ulster.ac.uk)

P. J. McCullagh  
School of Computing, Ulster University, Newtownabbey, Northern Ireland, UK  
e-mail: [pj.mccullagh@ulster.ac.uk](mailto:pj.mccullagh@ulster.ac.uk)

## 15.1 Self-quantification

Sensor-enriched environments and wearable technologies have the potential to collect data to quantify health and wellbeing. Many gigabytes per day, can be captured, often in a variety of heterogeneous formats [14] and a ‘Quantified-Self’ movement [25] has emerged. This has in turn necessitated research into the accessibility, extraction, mining and categorising of the resulting ‘big’ datasets. The areas of self-monitoring include physical activity, diet and nutrition, location, social interactions and visual interaction using photographs or videos [30].

With the ubiquitous nature of wearable sensors, ambient intelligence [6] and internet-connected devices [21], an Internet of Things (IoT) has emerged which has the ability to quantify many aspects of a person’s life. Devices such as pedometers, heart rate sensors, blood pressure cuffs and global positioning system (GPS) trackers allow people to record and analyse their lifestyle activities and aspects of their health profile. Behaviour change strategies often follow, promoting exercise, healthy diet, better sleep patterns in order to facilitate lifestyle changes [23]. Moreover, the data captured can be shared with healthcare professionals, or more commonly with friends and family via social media [18]. Paradoxically, users of technologies that facilitate self-quantification are more likely to be persons with an already healthy lifestyle. As such, device manufacturers target the youth and sporting markets. Nevertheless, the process of life-logging can have many benefits to the mature and older demographic, people with chronic diseases [3], memory impairments [10] and mental health conditions [30].

### 15.1.1 Life-Logging Technology

Life-logging is the process of capturing and recording data generated by behavioural activities. These activities include physical activity, eating and sleeping. Sensors specifically related to the logging of physical activity are now included with software apps in nearly all smartphones. Steve Mann has been at the forefront of technologies for capturing life-log data in the form of images [19, 20]. While self-quantification appeals to many and could promote a healthier lifestyle, the use of image recording in particular has opened an ethical debate regarding implications of capturing life-logs in public spaces [29]. The recording of images greatly increases the need for storage and can produce an uninteresting dataset that is labour intensive to analyse, with significant redundancy.

Using technology, it is possible to monitor almost every aspect of our daily lives, including where we go, with whom we interact, what we have eaten and what we see. The data set can also include sound, geo-location, accelerometer and orientation information and bio-signals such as temperature, heart rate, and oxygen saturation. By recording these data, it is possible to draw correlations between activities and behaviour. In order to facilitate behaviour change, feedback loops can be provided

both in real time and during review phases [23]. The use of wearable technologies is merely a facilitator to behaviour change. By creating feedback loops, that capture the user's attention at the moment they are most likely to take action, technologies can encourage sustained engagement [8].

Location and physical activity metrics are the most common parameters that have been captured by smartphones and their in-built global positioning system (GPS) sensors and accelerometers. Wearable life-logging devices that captured GPS location as well as first person 'point of view' images were pioneered by commercial devices such as SenseCam, Autographer and Narrative. Gurrin et al. [12] performed a comparison of the capabilities of the smartphone (as a life-logging device) with the SenseCam. Smartphones were given to 47 participants in order to record one day's worth of life-logs. These data were compared to a study by Doherty et al. [9] that used the SenseCam. The authors concluded that due to the rich user interface, additional sensors and wireless connectivity (Wi-Fi, GSM and Bluetooth), the smartphone had the capability to become a ubiquitous life-logging technology, which could be applied in the area of health research. Additional technology such as augmented reality glasses [11] provides overlays of digital content onto real-world objects and environments and this could increase the effectiveness of the feedback loop to promote behaviour change.

### ***15.1.2 Technology for Affective Computing***

Affective computing is the development of technology that can recognise, interpret, process and stimulate human feelings and emotions [17]. By incorporating multidisciplinary fields such as psychology, computer science and cognitive science, systems that can interpret the emotions of humans in order to control computing systems can be developed [24]. By having a computer system interpret the emotional state of the user, the services of the computer can adapt to the emotion and as such provide a much more pleasant user experience. The use of emotion measures can be investigated to determine the emotional relevance of life-logs.

Massachusetts Institute of Technology developed SmileTracker, which uses facial recognition to detect the user's expressions in order to track emotions when browsing the Internet. By collecting an image of the user's face and a screenshot of the current webpage, a record of happy browsing is generated for later reflection. The user's emotional state has been used to personalise user-generated content for sharing on social media. Shirokura et al. [27] used Galvanic Skin Response (GSR) sensors to collect the emotional state of a user while recording video content. Following data collection, visual effects may be added to the captured video, based on the user's mood during recording. This would then be shared and allow remote viewers to get a glimpse of the creator's emotional state. This system comprising a smartphone application and a wrist-worn GSR sensor was evaluated by two participants and demonstrated a 'proof of concept', indicating that physiological signals can be used to personalise video. This work demonstrates the potential role that affective computing

plays in personalisation of datasets, and this concept could be used as a measure of relevance when collecting life-log data. Irrelevant or boring sequences can be potentially dismissed at source, thereby reducing the data capture significantly.

Kelly and Jones [15] tested the hypothesis that it is possible to use physiological signals, specifically GSR in the identification of life-logged events for self-reflection. Three participants were asked to review life-logged data such as SMS messages and SenseCam images one month after collection. Results demonstrated that GSR could be used to extract important events from a large dataset. The affective diary study explored the emotional aspect of creating diaries for self-reflection [28]. By using a wearable armband equipped with GSR sensors, the life-log was augmented with the wearer's physiological signals. For feedback of the collected data, coloured 'blobs' were displayed on screen to show the wearer's emotional state when the life-logged data were displayed.

Some commercial devices allow developers access to the raw data that is collected from the sensor. These data can then be analysed, in real time on a smartphone allowing personalised and contextual feedback to be given. For example, Shimmer is an open platform for developers to collect kinematic and physiological data to be analysed offline or in real time (streamed via Bluetooth). Sensor options include accelerometer, gyroscope, electrocardiogram, GSR and electroencephalogram.

Not only are wearable technologies used to capture the emotions of users for affective computing, but research has been undertaken into the use of technologies to alter the user's mood. The Thync claims to use ultrasonic waves to stimulate specific neural pathways of the brain in order to alter the user's state of mind and energy levels [4, 31], thereby closing the affective computing loop.

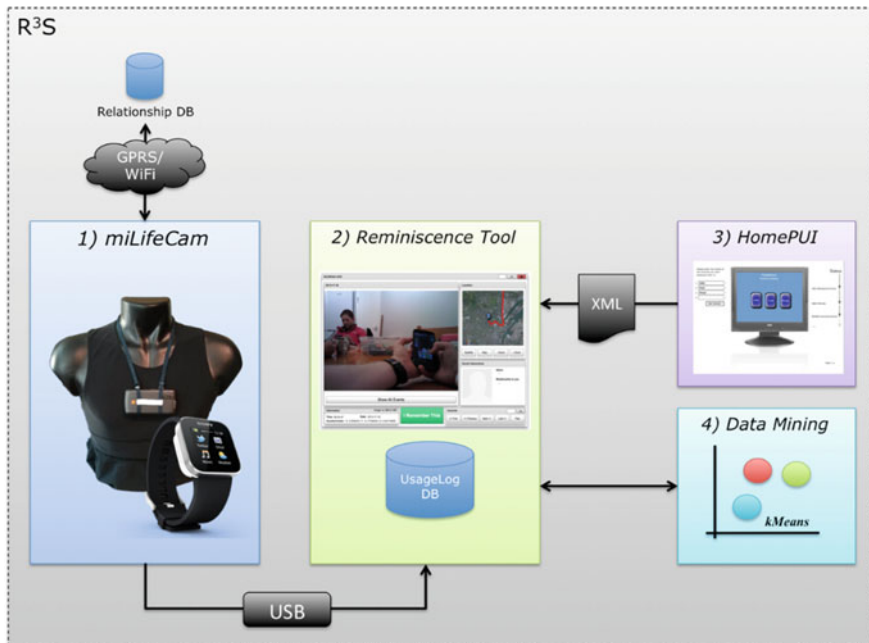
## 15.2 Methods

In our work, we have developed a wearable life-logging system, by integrating a smartphone, smartwatch and sensors. The system called R<sup>3</sup>S (Real-time social Reminder and Reminiscence System) [5] provides a suite of applications that facilitates the personalisation, capture and review of life-logging data during daily activities.

There are four components, see Fig. 15.1:

- *miLifeCam*, a wearable system to collect life-log data and identify known social contacts and provide real-time social interaction reminders;
- *Reminiscence Tool*, which facilitates the review and visualisation of the collected life-log data;
- *HomePUI* (Personalised User Interface), a tool for the personalisation of the look and feel of user interfaces;
- Geo-Data Mining, a tool to personalise the collected life-log data based on usage patterns collected by the *Reminiscence Tool*.





**Fig. 15.1** Architecture of the real-time social reminder and reminiscence system, detailing its four main components

This chapter focuses on the first two aspects. Six participants, three male and three female, were recruited from Ulster University’s School of Computing (academic and research staff and postgraduate students, 24–46 years of age, mean age 35.3). All but one of the participants had expert computing experience, with the remaining participant having a moderate level of expertise. Participants wore the smartphone located on their chest for a period of approximately two and a half hours and collected life-logs in the form of images, GPS location, accelerometer data and a record of social interactions. The Faculty of Computing and Engineering Research Ethics Filter Committee at Ulster University granted ethical approval for this study (2013 09 04 13.37). For privacy reasons, the participants were informed that they should either pause the recording, using the device’s hardware buttons, or remove the system when they were performing any sensitive tasks, e.g. using the bathroom. The *miLifeCam* component was investigated with regard to its appropriateness and acceptability as a wearable computing device. The reliability of the identification of known individuals, using body-worn Quick Response (QR) codes was evaluated, and information delivery to the smartwatch was assessed.

The system was designed to be wearable, comfortable and unobtrusive. To assess usability, the participants used the R<sup>3</sup>S ‘in the wild’, without it impacting significantly on their usual daily activities. To test the *miLifeCam*’s ability to recognise known social contacts, a predetermined interaction took place at least once during the

evaluation phase. Due to the battery drain on the device, caused by using the smart-phone's features/sensors (WiFi, Bluetooth, Camera, GPS and accelerometers.), the evaluations lasted a maximum of 2.5 h (The empirically measured duration of the battery life of the device with the developed software running).

Participants completed questionnaires on the usability and robustness of the developed system and based on QUEST 2.0 [7]. This evaluation was designed to ascertain the usability of the hardware components in addition to the collection of realistic life-log data. The evaluations investigated the effectiveness of life-logging as a means of (i) acquiring image data, (ii) acquiring meta data for subsequent summarisation, (iii) identifying known individuals and (iv) the overall system robustness.

### ***15.2.1 Life-Logging Data Capture***

Each participant was asked to wear the *miLifeCam* which was chest located on a lanyard. The *miLifeCam* component uses an Android smartphone and smartwatch in addition to a shirt embedded with textile electrodes and sensors. The solution facilitates recording daily activities, GPS location and heart rate, in addition to prompting participants when they interact with a known social contact. The application also captures accelerometer data that can determine the wearer's spatial orientation.

The system recorded GPS location (ascertained using GPS or assisted-GPS depending on location), accelerometer data, any social interactions and an image of what the participant was looking at. This data set was collected every 30 s. During the recording phase, each participant was approached by a researcher wearing a QR code badge with a unique identification number. This was to test the speed with which the system would identify a social interaction and provide feedback via the smartwatch. This identification technique was designed for the proof of concept phase and can be updated with facial recognition software (for a pre-determined group of contacts).

### ***15.2.2 Summation and Personalisation of Life-Log Dataset***

Recording an image and metadata every 30 s would result in 1440 images and corresponding data points (timestamps, GPS, etc.) over the course of a 12-h period. Over a week, this would result in over 10,000 images, each one with an average file size of 250 kb; complete with metadata this totals approximately 2.7 GB. While the technical feasibility for the capture is not in question, the analysis of such a large dataset in order to be categorised and personalised for review presents a significant challenge. The data captured, in the form of images and metadata, during this evaluation was used to evaluate the four summation and categorisation methods.

- (i) Self-reporting of interesting/relevant images. This was undertaken by the participant who collected the data. It was used to reduce the life-log dataset size and generate more personalised life-log events.
- (ii) The metadata captured during life-log recording was utilised to cluster events in the dataset based on GPS location. Using a k-means algorithm, the metadata was parsed and clusters were generated based on 15-min timeframes, i.e. 30 life-log images.
- (iii) An algorithm identified images that contained a face.
- (iv) GSR data, collected during review phase of the life-log, was utilised to identify any physiological changes during review. This may indicate higher relevance.

## 15.3 Data Reduction Processes

### 15.3.1 Self-reporting Phase

Following the data collection, participants were asked to upload their recorded data to the *Reminiscence Tool* for review. Each participant completed this task on the same day as the recording. The participants were then asked to interact with the reminiscence software with no previous interaction or instruction. The researcher informed the participants of the controls available to them, namely the ability to move between images one at a time, jump to the first or last images or review the images sequentially using the 'Play' button. User interface interactions were captured and stored in the *Reminiscence Tool's* SQLite database. The participants were asked to press the '*I Remember This*' button when they came across an image that stimulated some feeling or memory. These 'flags' were also stored in the database. During the interaction with the *Reminiscence Tool*, the participant's eye movements were tracked using a Tobii eye-tracking system to assess screen interaction.

By monitoring how the user navigates through the life-log data, it is possible to personalise existing and future events based on these patterns. Parameters such as the time spent looking at a particular image and the flagging of an image as prompting some form of reminiscence can be used to assess important data. The *Reminiscence Tool* monitors usage parameters and stores:

- The date/time the data were collected
- The image being viewed
- Any associated button presses
- Reminiscence flag as identified by participant
- Timestamp of all interactions.

This usage information is collected to enable the system to personalise the presentation and organisation of recorded data based on user preferences. The first stage was to identify all images that were flagged using the "*I Remember This*" button. For these images, the preceding and succeeding two images were automatically selected

as a sub-event. In addition to these user analytics, the GPS coordinates of the images in this sub-event allow additional images to be queried for distance to that location, thereby creating a location event. This assumes the reason the user pressed the “*I Remember This*” button was due to the image’s location and not specific content (refer to Fig. 15.2, Reminiscence AOI). It is of course possible that the people in the image meant significantly more to the user than the location in which the interaction took place.

By preforming a query on the SQLite database of the *Reminiscence Tool*, it is possible to reduce the current datasets (average 313 images) to a smaller and more relevant dataset. The reduced event size is the number of reminiscence events times (two previous images, the flagged image and two images after the flagged image). Five images give a snapshot of 2.5 min of life-log time. In the event that another reminiscence flag is within a previous window, the following two images of the second flag are added to a larger window. While this method can significantly reduce the number of images that are displayed, it requires an initial viewing in order to flag the images as triggering some form of reminiscence. Nevertheless, the same process can be applied automatically to images that have social interaction metadata, as the *miLifeCam* component flags the images as having a social interaction event. Using the self-report method outlined above, the life-log dataset can be reduced by an average of 86% from an average dataset size of 313 images (see Table 15.1). This reduces review time from an average 10 min 26 s to 1 min 28 s, assuming images are replayed at two-second intervals.

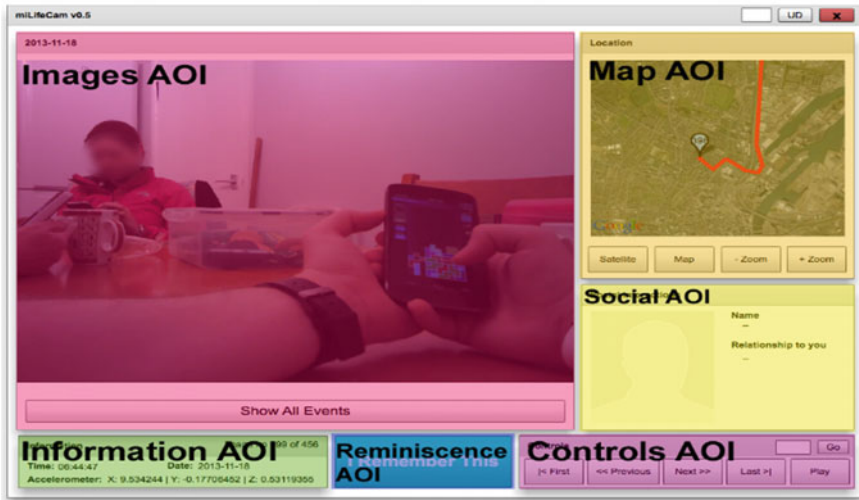


Fig. 15.2 Reminiscence tool’s interface with areas of interest highlighted: (Clockwise from top-left), Images, Map, Social, Controls, Reminiscence and Information Areas of Interest (AOI)

**Table 15.1** Percentage reduction of each participant's life-log dataset using each of the presented methods

| Participant | Self-report | Geo-mining | Face detect | GSR   |
|-------------|-------------|------------|-------------|-------|
| 1           | 75.23       | –          | 79.8        | –     |
| 2           | 93.13       | –          | 83.9        | –     |
| 3           | 84.20       | 90.65      | 90.97       | 72.76 |
| 4           | 91.70       | 67.39      | 88          | 80.62 |
| 5           | 89.86       | 95.23      | 76.7        | 85.38 |
| 6           | 84.60       | –          | 61.4        | 83.70 |
| Averages    | 86.45       | 84.42      | 80.12       | 80.61 |

Three of the six participants were unable to record GPS location data used for the geo-mining method and two were unavailable to take part in the GSR data collection

### 15.3.2 Geo-Data Mining

By using geo-data mining techniques, it may be possible to extract only relevant sub-events from the larger event set. This could be achieved by clustering the images based on the GPS location. GPS location contains two coordinates (latitude and longitude) stored in a two-dimensional vector. As result, a k-means clustering algorithm was applied for further data reduction. The total number of images in the dataset was divided by 30, as each image was captured at 30-s intervals. This means that each possible cluster spans a 15-min period, sufficient time to move from one location to another ( $30 * 30 / 60 = 15$ ). Kikhia et al. [16] have taken a similar approach to the structuring of life-log data for the purposes of reminiscence. GPS locations of all the participant's data were clustered using WEKA's SimplekMeans Algorithm [13].

Images were categorised based on physical distance, calculated by GPS coordinates, from a cluster (e.g. images within an 11-m radius of the GPS centroid location). This was calculated by querying that dataset for all locations within a range of the generated centroids, i.e. by modifying the fourth decimal places of the latitude and longitude coordinates [32]. This technique poses a problem if a user remained sedentary for extended periods of time, for example, if the user remained at home for two hours; all 240 images captured would be included in the generated event. Using this method, the dataset can be reduced on average by 84%. (Note due to some technical difficulties the data set only includes data for three participants.)

### 15.3.3 Face Detection

A feature of the R<sup>3</sup>S system is the automatic detection of social interactions with known individuals using QR codes. As each social contact is required to be wearing a QR code in order for the system to recognise them, there is the potential for other social interactions, with persons not wearing a QR code, to be missed. Face detection

may be used as a means to identify interactions with known and unknown persons, and as a means to reduce the dataset size.

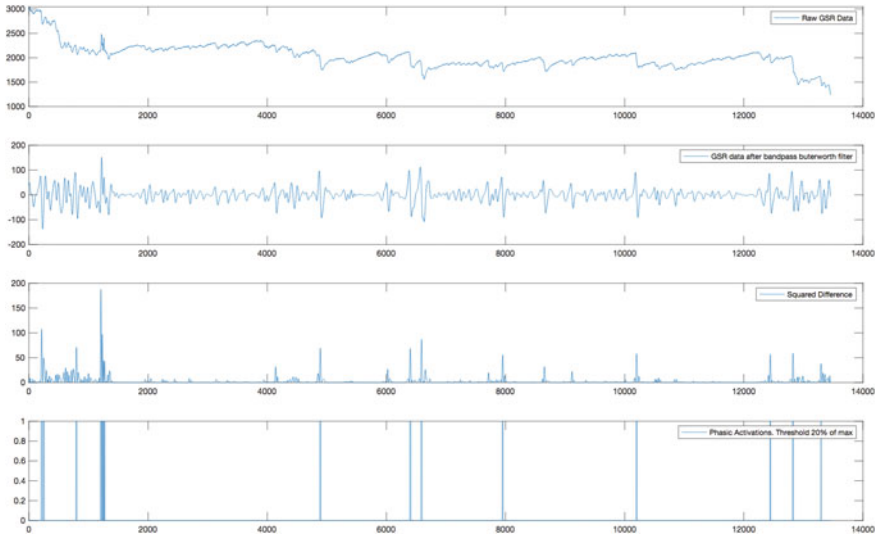
Each participant dataset was manually reviewed to ascertain the number of images in which a face, of any person, appeared. Only images that were face on, i.e. displayed two eyes, nose and mouth were considered. Manual collection was compared to an automatic face detection algorithm developed by Google for the Android platform. This algorithm looks for specific facial landmarks such as eyes, nose, cheeks and mouth in order to detect a 'face'. By default, the algorithm only detects a face when it has a confidence threshold of 0.4. In order to reduce the overall life-log a pruning method of the preceding and succeeding two images, similar to the self-report method above was used.

The performance of the automatic face detection algorithm was comparable with that of the manual detection. Nevertheless, in some cases, the differences were significant. Further analysis show that in some cases, faces were detected where none existed or faces on posters or screens were identified. Using the face detection method, the datasets were reduced, on average, by 80%. This automated process can be completed in a matter of seconds, depending on the dataset size, and the results presented to the user in a reduced 'social' dataset. As the algorithm used was native to the Android platform, it would be feasible to implement this algorithm into the *miLifeCam* component to capture social interactions at source and thereby reduce the amount of data captured. It is important to note that this method only 'detects' and does not 'recognise' faces.

### ***15.3.4 Physiological Recording Using Galvanic Skin Response***

Participants were asked to wear a GSR sensor on their left hand, then review the life-log images they previously collected using the *Reminiscence Tool*, six months after initial collection. The GSR was sampled at 20 Hz and stored in comma separated values (CSV) format for processing. GSR was recorded in order to ascertain if specific images presented to the participant stimulate a physiological response. Thus, it could potentially act as a trigger for life-log data capture. GSR data were analysed to detect any changes in physiological response linked with the visual stimuli presented to the participant. The raw GSR data were subjected to a four-part process to identify phasic changes for use in the reduction of the life-log datasets, as shown in Fig. 15.3. Unlike skin conductance tonic levels, which is a slow variation in skin conductance over tens of seconds to minutes, the phasic response rides on top of the tonic changes and occurs between 1 and 5 s after the stimuli. A derivative approach was developed to the detection of rapid changes in the signal. This approach is inspired by a technique widely cited in the analysis of other biosignals [22].

Initial signal filtering was achieved with the application of a fourth-order Butterworth bandpass filter with upper and lower cut off frequencies of 0.1 and 0.15 Hz,



**Fig. 15.3** Visualisation of participant six's GSR data. Raw GSR data is presented (top), followed by the phases of signal processing: Bandpass Butterworth filter, Squared Differential on the filtered data, and Phasic Activation (Threshold, 20% of max amplitude)

respectively. This filter was applied to the raw GSR signal to remove the baseline drift and high-frequency noise [1, 26]. Zero-phase filtering was employed to eliminate any phase distortion introduced by the nonlinear Butterworth response. Identifying the stimulus that the participant was looking at during collection and synchronisation of the signal timestamp is important. Following the initial smoothing of the data, the first derivative of the signal was obtained. In order to further emphasise the more significant changes in the signal a nonlinear scaling was also performed. In this step, the first derivative values were squared.

The final phase of the data analysis is the identification of a phasic activation. The max value of the integral was ascertained and an arbitrary threshold of 20% of the maximum values was applied to the results. As a result, any phasic activation that exceeds the threshold is identified. By identifying occurrences of these activations and synchronising these with the timestamps, the visual stimuli can be identified. By flagging the life-log image at this point, coupled with the previous and following life-log images, the larger life-log dataset can be reduced to a series of 'Phasic-Events' (PEs), visualised in Fig. 15.3. In some instances, the phasic activation (PA) identified occurred near a previous PA, as a result when two PAs are in this configuration the preceding image of the first PA is flagged, in addition to both PAs and the succeeding images of the latter PA is also flagged. Depending on the number of PAs in close proximity, this would result in at least five images occurring in that PE. Using the PA method, the datasets were reduced, on average, by 81%.

Some of the identified PAs during recording occurred during an initial calibration phase where generic images were displayed as opposed to the participant's life-log

data. Following the analysis of the GSR data from each participant and the synchronisation with the visual life-log stimuli several PEs were generated for reminiscence. While we can match the visual stimulus and GSR during the reminiscence evaluation, it is not conclusive that the images presented were the main cause of the phasic change in the GSR. The addition of eye-tracking technology during the experiment would go some way to answering this question more reliably. Nevertheless, a method of the objective identification of phasic changes GSR signals and their use in the reduction of larger life-log datasets is presented and shown to be feasible.

## 15.4 Discussion

The four data reduction methods presented have been shown to reduce life-log datasets, in the form of images and metadata, by approximately 80%, as presented in Table 15.1. The methods presented can be scaled up to larger datasets without significant additional data processing. Nevertheless, the Self-Report and GSR methods both require the user to review the dataset and provide input, in the form of user input or subconscious physiological information.

The face detection and geo-mining methods presented can be performed without the need of any user input and thus create a reduced dataset that can be reviewed by the user. Nevertheless, the automated reduction could potentially remove images from the dataset that ‘could’ elicit a physiological or self-reported response.

The *miLifeCam* system has the capability to identify and report in real-time social interactions with known individuals, using QR codes. This metadata could be used to categorise life-logged events and provides a mechanism for identifying social interaction, which could be appropriate to vulnerable cohorts of users.

Using the four methods presented above, it is possible to reduce the dataset size of life-logged data significantly, with an average of 86% reduction using the user self-report methodology and an 84% reduction using the geo-mining approach and 80% when using the face detection method. The final data reduction using GSR data results in an average 81% reduction. Because of these reductions, excessive amounts of collected data may no longer be required.

## 15.5 Conclusions, Limitations and Future Work

We presented four methods for the summation and personalisation of large life-logged datasets. Each of the methods presents a solution to the ‘big data’ challenge of life-logs, but there is a real need for automation and data reduction at source. Self-reporting of ‘reminiscence flags’ ensures that the events flagged have been consciously selected by the user and thus will likely have a benefit when used for personalisation. Nevertheless, this method, due to requiring the user to manually review all the data, is time-consuming.



The geo-data mining methodology used metadata captured during recording and thus does not require user input to generate smaller datasets. Using the GPS location of the recorded life-log facilitates the clustering of ‘events’ to reduce the dataset size. A challenge for this approach is posed when the user does not move location during significant events. For example, a user who stays at home may not physically move location during the collection phase, however, may interact with multiple visitors. These individual interactions would not be detected by the geo-data mining method but could be signalled by additional metadata labels. The capability of the *miLifeCam* system to identify and log social interactions in real time would allow the geo-data mining method to be modified to include social interactions as a means to segment the larger datasets.

The face detection method was another automated method that reduced the life-log dataset size of the participants. This method required no user interaction for reduction, nevertheless, there are some flaws such as the identification of false positives and artificial interactions. Due to the ability to detect faces, this method does not distinguish between an actual social interaction and that of being in a public space or looking at a television or poster.

The physiological method investigated the use of GSR sensors to ascertain if there is a physiological response from participants reviewing life-logged data collected over six months prior and if that physiological data could be used to personalise the life-logged data. The reminiscence of life-logged images after six months did have some quantifiable effect on the participants evaluated. However, no particular event type can be definitively identified to cause a consistent GSR response. Of the GSR events identified, in the highest number, the participants were shown to be looking at images depicting them as stationary, i.e. not with a person or moving between environments/rooms/places. Social interactions were the second-most common event type.

There is nothing from the findings that would suggest a specific activity or location prompts any significant and uniform change in GSR that could be used to ‘tag’ a particular life-log image in order to personalise the dataset. Nevertheless, this was a small exploratory study with the view to using GSR data as a means of triggering life-log capture, and hence reducing data size. Additional stimuli during the life-log recording, such as smells, sounds and even atmospheric and environmental measures, could trigger a more consistent GSR response than reminiscing with images alone, in silence [2].

Additional features contained on the smartphone, such as accelerometer and microphone, have the potential to be used to personalise the life-log. Using accelerometers and gyroscopes, physical activity could be recognised. Simple activities, such as walking and sitting, can be used to personalise the dataset. Nevertheless, a high sampling frequency would be required to ascertain these activities. Capturing and processing sound from the user microphone and processing it for keywords could also be a viable trigger for data capture/personalisation. For example, if the user was in conversation, keywords such as names, places and objects could be identified and used to flag images in the dataset, in a similar fashion to the self-report method.

For life-logging to become viable, automated data reduction at source is required. The data reduction method chosen may well depend on the weather monitoring and quantifying physical activity, sedentary behaviour, social interaction or assessing nutrition is the primary concern. Technology solutions are available in the form of geo-location, identification of social contacts and assessment of significance with GSR (or possibly heart rate variability). These can be deployed with a smartphone and accessories (e.g. body-worn camera, smartwatch). In addition, the software must be embedded and personalised to the user and the desired application, but again this could be readily achieved using the computation resource available on a smartphone. Self-quantification is not without its problems, particularly ‘in the wild’. We encountered issues with missing data (Table 15.1), and there are significant ethical and acceptance issues to be addressed for the wearer and her/his daily contacts.

The technology will continue to be adopted by niche applications, e.g. sports enthusiasts and security personnel, who value a record of their daily interaction. However, the approach described in this chapter could be tuned to meet the needs of a vulnerable person (as a tool for caregivers); for example, it could be used as an assistive technology to infer nutritional intake and social contact by coupling the timestamps of the recorded images, the location in which they are taken and the contents of the image.

## References

1. Bach DR, Friston KJ, Dolan RJ (2010) Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations. *Int J Psychophysiol* 76(1):52–55
2. Bakker J, Pechenizkiy M, Sidorova N (2011) What’s your current stress level? Detection of stress patterns from GSR sensor data. In: 2011 IEEE 11th international conference on data mining workshops (ICDMW), pp 573–580
3. Barrett MA, Humblet O, Hiatt RA, Adler NE (2013) Big data and disease prevention: from quantified self to quantified communities. *Big Data* 1(3):168–175
4. Boasso AM, Mortimore H, Silva R, Aven L, Tyler WJ (2016) Transdermal electrical neuromodulation of the trigeminal sensory nuclear complex improves sleep quality and mood. [bioRxiv](#)
5. Burns W, Nugent C, McCullagh P, Zheng H (2014) Design and evaluation of a smartphone based wearable life-logging and social interaction system. In: *Ubiquitous Computing and Ambient Intelligence*, pp 179–186
6. Cook DJ, Augusto JC, Jakkula VR (2009) Ambient intelligence: technologies, applications, and opportunities. *Pervasive Mob. Comput.* 5(4):277–298
7. Demers L, Weiss-Lambrou R, Ska B (2002) The Quebec user evaluation of satisfaction with assistive technology (QUEST 2.0): an overview and recent progress. *Technol Disabil* 14(3):101–105
8. Direito A, Dale LP, Shields E, Dobson R, Whittaker R, Maddison R (2014) Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? *BMC Public Health* 14(1):1
9. Doherty AR et al (2011) Passively recognising human activities through lifelogging. *Comput Human Behav* 27(5):1948–1958
10. Elsdon C, Kirk DS (2014) A quantified past: remembering with personal informatics. In: *Proceedings of the 2014 companion publication on designing interactive systems*, pp 45–48

11. Etherington D (2017) Google Glass is back with hardware focused on the enterprise, TechCrunch.com
12. Gurrin C et al (2013) The smartphone as a platform for wearable cameras in health research. *Am J Prev Med* 44(3):308–313
13. Hall M, Frank E, Holmes G (2009) The WEKA data mining software: an update. *ACM SIGKDD* 11(1):10–18
14. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of ‘big data’ on cloud computing: review and open research issues. *Inf Syst* 47:98–115
15. Kelly L, Jones G (2010) An exploration of the utility of GSR in locating events from personal lifelogs for reflection. In: *iHCI 2010—4th Irish human computer interaction conference*, pp 82–85
16. Kikhia B, Boytsov A, Hallberg J, Sani H, Jonsson H, Synnes K (2014) Structuring and presenting lifelogs based on location data. In: *Pervasive computing paradigms for mental health*, pp 133–144
17. Luo J (2012) *Affective computing and intelligent interaction*, vol. 137. Springer
18. Lupton D (2013) Quantifying the body: monitoring and measuring health in the age of mHealth technologies. *Crit Public Health* 23(4):393–403
19. Mann S, Niedzviecki H (2001) *Cyborg: digital destiny and human possibility in the age of the wearable computer*. Doubleday Canada
20. Mann S, Nolan J, Wellman B (2002) Sousveillance: inventing and using wearable computing devices for data collection in surveillance environments. *Surveill Soc* 1(3):331–355
21. Miorandi D, Sicari S, De Pellegrini F, Chlamtac I (2012) Internet of things: vision, applications and research challenges. *Ad Hoc Netw* 10(7):1497–1516
22. Pan J, Tompkins WJ (1985) A real-time QRS detection algorithm. *IEEE Trans Biomed Eng* 3:230–236
23. Patel MS, Asch DA, Volpp KG (2015) Wearable devices as facilitators, not drivers, of health behavior change. *JAMA*
24. Picard RW (2000) *Affective computing*. MIT press
25. Piwek L, Ellis D, Andrews A, Joinson A (2016) The rise of consumer health wearables: promises and barriers. *PLoS Med.* 13(2):e1001953
26. Schumm J, Bachlin M, Setz C, Arnrich B, Roggen D, Troster G (2008) Effect of movements on the electrodermal response after a startle event. In: *Second international conference on pervasive computing technologies for healthcare, PervasiveHealth 2008*, pp 315–318
27. Shirokura T, Munekata N, Ono T (2013) AffectiView: mobile video camera application using physiological data. In: *Proceedings of the 12th international conference on mobile and ubiquitous multimedia*, pp 1–4
28. Ståhl A, Höök K, Svensson M, Taylor A, Combetto M (2009) Experiencing the affective diary. *Pers Ubiquitous Comput* 13(5):365–378
29. Stenovc T, Mann S (2012) Inventor, Allegedly Attacked At Paris McDonald’s For Wearing Digital Eye Glass, *Huffington Post*. [Online]. Available [http://www.huffingtonpost.com/2012/07/17/steve-mann-attacked-paris-mcdonalds-digital-eye-glass-photos\\_n\\_1680263.html](http://www.huffingtonpost.com/2012/07/17/steve-mann-attacked-paris-mcdonalds-digital-eye-glass-photos_n_1680263.html)
30. Swan M (2013) The quantified self: fundamental disruption in big data science and biological discovery. *Big Data* 1(2):85–99
31. Tyler WJ et al (2015) Transdermal neuromodulation of noradrenergic activity suppresses psychophysiological and biochemical stress responses in humans. *Sci Rep* 5:13865
32. Zandbergen PA (2009) Accuracy of iPhone locations: a comparison of assisted GPS, WiFi and cellular positioning. *Trans GIS* 13(s1):5–25

# Chapter 16

## Privacy-Enabled Smart Home Framework with Voice Assistant



Deepika Singh, Ismini Psychoula, Erinc Merdivan, Johannes Kropf, Sten Hanke, Emanuel Sandner, Liming Chen and Andreas Holzinger

**Abstract** Smart home environment plays a prominent role in improving the quality of life of the residents by enabling home automation, health care and safety through various Internet of Things (IoT) devices. However, a large amount of data generated by sensors in a smart home environment heighten security and privacy concerns among potential users. Some of the data can be sensitive as it contains information about users' private activities, location, behavioural patterns and health status. Other concerns of the users are towards the distribution and sharing of data to third parties. In this chapter, we propose privacy-enabled smart home framework consisting of three major components: activity recognition and occupancy detection, privacy-preserving data management and voice assistant. The proposed platform includes unobtrusive sensors for multiple occupancy detection and activity recognition. The privacy-enabled voice assistant performs interaction with smart home. We also present a detailed description of system architecture with service middleware.

**Keywords** Smart home · Activity recognition · Occupancy detection · Privacy-preserving data management · Dialogue manager

---

D. Singh (✉) · E. Merdivan · J. Kropf · E. Sandner  
AIT Austrian Institute of Technology, Wiener Neustadt, Austria  
e-mail: [deepika.singh@ait.ac.at](mailto:deepika.singh@ait.ac.at)

I. Psychoula · L. Chen  
School of Computer Science and Informatics, De Montfort University, Leicester, UK

E. Merdivan  
CentraleSupélec, Metz, France

S. Hanke  
FH JOANNEUM Gesellschaft mbH, Graz, Austria

D. Singh · A. Holzinger  
Holzinger Group, HCI-KDD, Institute for Medical Informatics/Statistics, Medical University  
Graz, Graz, Austria

## 16.1 Introduction

In recent years, smart home technology along with Internet of Things (IoT) devices has gained a lot of attention due to its various applications. It has experienced rapidly growing presence in the households of end-users. A smart home comprises of sensors such as motion sensors, thermostat, passive infrared sensors, energy tracking switches, smart lights, shades, cameras and voice assistants, which communicate with each other and collect data to monitor user activities. The data collected from these embedded sensors and smart devices provide numerous services to the residents such as safety and guidance features by user behaviour monitoring, activity recognition and fall detection; home automation by controlling lights, doors, windows, temperature and energy consumption; and security with alarms, lock/unlock of doors and monitoring of outsiders in the absence of resident.

Various devices such as tablets, personal computers and smartphones are being used for communication and interaction with users in a smart home environment. Recent works have shifted towards dialogue systems in the form of voice assistant. The integration of the dialogue system in a smart home can provide a natural and convenient way of interaction with a user. Dialogue systems are categorized mainly into two groups [5] as task-oriented systems and non-task-oriented systems (also known as chatbots). The aim of task-oriented systems is to assist users to complete specific tasks by understanding the inputs from the user such as restaurant bookings, ticket bookings or information searching. Non-task-oriented dialogue systems can communicate with humans on open domains, and thus, they are preferred in real-world applications. In a smart home environment, task-oriented dialogue systems can assist users in performing various tasks depending on their needs. For example, a medication reminder can check if a user has taken medicine as prescribed by their doctors, and further, recommend exercises or any other physical activity when sensors in the home detect that a user is in a sedentary state for a considerably longer time. It could also make an appointment with a doctor on behalf of the user. A dialogue system can also be customized according to the user's preferences.

A user's interaction with a dialogue system and sensors installed in the home environment generate a large amount of data. Such data usually contain private and sensitive information such as users' profile, locations, pictures, daily activities and network access information, which leads to privacy, and security concerns for the users. These raised concerns include "who manage and regulate these devices?", "where the data is stored?" and "who has access to the data?" [41], as well as how do machine learning systems reach the decisions they do and how fair and non-biased are the algorithms used in them? [37, 38].

Previous studies have introduced various smart home frameworks with network-level security and privacy in IoT devices. A three-party architecture with Security Management Provider (SMP) has been proposed, which identifies and blocks the threats and attacks at the network level [35]. Passive network adversary can infer private activities of a user from smart home traffic rates even when devices use encryption, this problem has been discussed in [3] by introducing stochastic traffic

padding technique. Different frameworks have been designed for user activity recognition [15, 21], occupancy detection [40], adaptive user interaction [17] and dialogue systems [24] in smart home environments. Most of the existing frameworks focus only on one particular component of the system, for example, only activity recognition module, adaptive interface or privacy in smart home IoT devices. Therefore, in this work we presented a framework, which combines three major components of the smart home system, i.e. user occupancy detection, privacy-preserving data management and dialogue system for user interaction. The proposed framework focuses on the development of technological solutions, which can monitor the users and environment while keeping their privacy intact.

The remainder of this chapter is organized as follows: the related work of each component of the framework is presented in Sect. 16.2. Section 16.3 gives an overview of the proposed framework. System architecture of the proposed framework with a detailed description of individual component is explained in Sect. 16.4. Section 16.5 presents the two different scenarios based on the proposed framework. The last section includes the conclusion and future work.

## 16.2 Related Work

In this section, we review related work focusing on the three major components of the framework, which are occupancy detection and activity recognition, privacy preservation in data management and dialogue systems.

### 16.2.1 *Activity Recognition and Occupancy Detection Module*

Human activity recognition has an important role in a smart home system by learning high-level knowledge about the user's daily living activities from sensor data. The monitoring systems can be categorized mainly into three categories: (1) camera-based, (2) wearable devices and (3) binary and continuous sensors. Camera-based monitoring is not preferred by the users as it is considered privacy invasive. Wearable devices provide good accuracy in the case of personalized systems and do not raise as many privacy concerns, and however, such systems are not practical to use while monitoring long-term activities. As a result, the most preferred solution for activity recognition is using unobtrusive sensors in a smart home. Early work on activity recognition have used data mining methods and knowledge-driven approaches, e.g. Chen et al. [6], Okeyo et al. [23]. In the recent work, deep learning approaches have contributed tremendously towards activity recognition systems [32, 33] and tend to overcome the limitations of conventional pattern recognition approaches. In deep neural networks, the features can be learned automatically instead of using manually

handcrafted features. The deep generative networks can also make use of unlabelled data for model training as in activity recognition systems; it is not feasible to have labelled data most of the time [37, 38].

Occupancy detection and estimation plays a major role in reducing energy consumption by controlling heating, ventilation and air conditioning (HVAC) and lighting systems in the smart environment. It also helps in the detection of intruders, abnormal events such as falls and monitoring activities of multiple residents inside a home environment. Different methodologies on occupancy estimation have been presented and discussed which include occupancy detection using the camera, passive infrared sensor (PIR), ultrasonic sensor, radio frequency signals (RFID), fusion of sensors and using wireless local area network (WLAN), Bluetooth and Wi-fi [2]. Camera-based occupancy estimation is accurate in comparison with other methods but due to privacy concerns, they are not preferable in real-time applications. PIR and ultrasonic sensors are economical but can only detect the presence/absence of an occupant as it produces a binary output. Use of RFID tags is not feasible in real-life situations as occupants have to carry RFID all the time and it has privacy issues as well. Occupancy detection through WLAN, Wi-fi and Bluetooth technologies is not applicable in large buildings due to its high positive/negative detection of occupants. Use of multiple sensors such as CO<sub>2</sub>, PIR, temperature, humidity, light and motion sensors for occupancy detection provides accurate results and can be easily applied in real-time applications. Therefore, we prefer data generated from fusion of multiple sensors for occupancy detection in the proposed smart home framework.

## ***16.2.2 Explainable AI in Smart Home Systems***

As can be seen in previous sections, the success of machine learning in activity recognition and advancement of smart home technology has been undeniable. However, there are still concerns about adopting these learning techniques in practice. One of the reasons is that with these techniques it is quite difficult to understand what goes on inside an algorithm and why it gave the result it did. This difficulty of explaining the reasoning behind the generated result makes the decisions provided by these systems hard to be trusted by the end-users. A lot of recent work has been dedicated to making machine learning models explainable [12, 13]. There are two main aims of work on interpretability in the literature: transparency and post hoc interpretation [1]. Transparent design reveals how a model functions, while a Post hoc Explanation explains why a black-box model behaved that way [19].

Explainability enables the privacy risk discussions that need to happen when decisions are being made in the development and design stages of a machine learning model. When the decisions that the model makes will affect individuals based just on the model's output, there is the possibility for bias and wrong decision-making. Many machine learning models include multi-layer neural networks, which follow an internal process in which outcomes may not be able to be understood in a mathematical way even by experts in the field. Due to this, multi-layer features, model performs

better. This is often referred to as the accuracy vs. interpretability trade-off which is also known as the “black-box” problem. Some of the solutions to this problem are following the steps of the algorithm and describe with more details about what the model is doing. Another way is to perform risk management assessments that evaluate the model, the data and the output and the risk to the individual if the decision reached is not correct which will also affect the privacy of the individual. Bias can often occur based on the selection process of a model’s features, and the weighting their assigned, during the implementation of the system. The decisions made when defining categories and establishing the relationships between them affect both the accuracy of the model and the outputs that will be produced; especially when the systems are making recommendations that affect individuals.

According to the General Data Protection Regulation [11], one of the requirements is the “right of explanation” which is a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place. Explainable AI will assist in making the decisions of machine learning algorithms more understandable and in this way help make the privacy implications of data sharing clearer to the end-users. Explainable AI is essential to build trust and increase the acceptance of smart home systems. Any inherent bias either in the data sets, in the weighting of various features or choosing which ones to use, has real-world results that must be fair and explainable by the system.

### ***16.2.3 Privacy in Data Management***

Generally speaking, data collected in smart home environments usually go through the following steps: first data collection, followed by data aggregation and finally data mining and data analytics. Although these steps enable a lot of services to be provided to the smart home users, they pose a lot of privacy challenges. For example, in the smart home environment if an adversary obtains data about the occupancy of the residents, he can infer the pattern of when users are inside or out of the house which could lead to theft or other damage to the users. Privacy-preserving mechanisms are necessary in these scenarios to protect the sensitive data and the privacy of the users.

Existing privacy-preserving mechanisms include anonymity, encryption, perturbation and privacy-preserving data analytics techniques such as differential privacy [9] and homomorphic encryption [22]. Recent techniques make use of deep learning mechanisms to provide privacy-preserving probabilistic inference [25], privacy-preserving speaker identification [26] and computing on encrypted data [4, 39]. While the previous work on privacy-enabled frameworks has focused on issues like assisting users with mobile app permissions [16], protecting user location data [14] and privacy-aware video streaming [8], the concerns and preferences of the users are not always taken into account.

Most of the time users realize what they want from a specific technology only after they have started using it. In most privacy surveys, the users make comments



like the following: “I want to decide what types of data a service or application gets access to” and “I should get to decide who my data is shared with” [28]. Users have significant reservations about continuous ongoing monitoring [34] in which sensors collect measurements and can store and analyse all the sensor data in the cloud. Users prefer to have control over how their data are used. For example, if data from door sensors and light sensors are being combined to detect room occupancy, the users might want to be aware that there is this possibility of detection since most sensors do not appear too intrusive on their own, but when combined they can reveal a lot of information. Also, users may wish to give access to their data, but may prefer to do so anonymously, for example, as medical data for research. So it is essential for privacy-enabled frameworks to support anonymization. However, most of the time users often find it difficult to understand privacy controls. That is why privacy-enabled voice assistant can be beneficial in helping users manage their privacy preferences and offer them more control over which data they share and at what granularity. In this work, we extend beyond existing privacy preservation techniques and propose a framework that allows users to control the flow of information in a privacy-preserving manner.

#### ***16.2.4 Dialogue Systems***

Lately, there has been tremendous growth in embodied conversational agents (ECAs) using natural language processing (NLP) techniques for developing intelligent dialogue components. NLP-based systems have gained a lot of interest in human-machine interactions for multi-modal interfaces. They are preferred widely for natural and spoken language understanding (NLU or SLU), natural language generation (NLG) and dialogue processing (DP) tasks.

A complete dialogue solution consists of various components which include automatic speech recognition (ASR), natural language interpreter, dialogue state tracker (DST), dialogue response selection (DRS) and text-to-speech (TTS) component [31], where each component has a specific task associated with it. Existing work on training dialogue managers includes rule-based methods; sequence-to-sequence methods; reinforcement learning (RL)-based methods; and hierarchical reinforcement learning (HRL) methods. In rule-based methods, a set of rules are defined by humans to train dialogue managers which make these systems robust and easy to use. However, it is not possible to use such systems for real-life applications. Sequence-to-sequence methods are used to transform a given sequence from one domain to another and are widely used in translation tasks such as English to French and performed well with less or none natural processing of sentences [36].

Reinforcement learning and hierarchical reinforcement learning have gained a lot of popularity due to their astonishing results in certain gaming tasks such as Atari, chess and GO [20] which outperform the human score but require a large amount of training data which are impractical to collect and simulate due to complexity and variety in human dialogues. HRL-based methods on dialogue management

outperform the standard RL methods, but there are still unresolved issues. Namely, those issues are in handling deeper hierarchies; designing a reward for lower level hierarchies; and also automatically dividing the complex task into simpler sub-tasks which are possible with deeper hierarchies than few levels of hierarchy.

### 16.3 Proposed Framework

This section presents different modules of the smart home framework. An overview of the proposed smart sensing framework is presented in Fig. 16.1. The hardware includes an integrated sensor network with devices that can monitor multiple occupants and the home environment, a gateway and local storage device to collect the data, along with a smart interactive user interface that enables the communication with the user and the controlling of the smart home. In the following subsections, the main elements of the proposed framework are described.

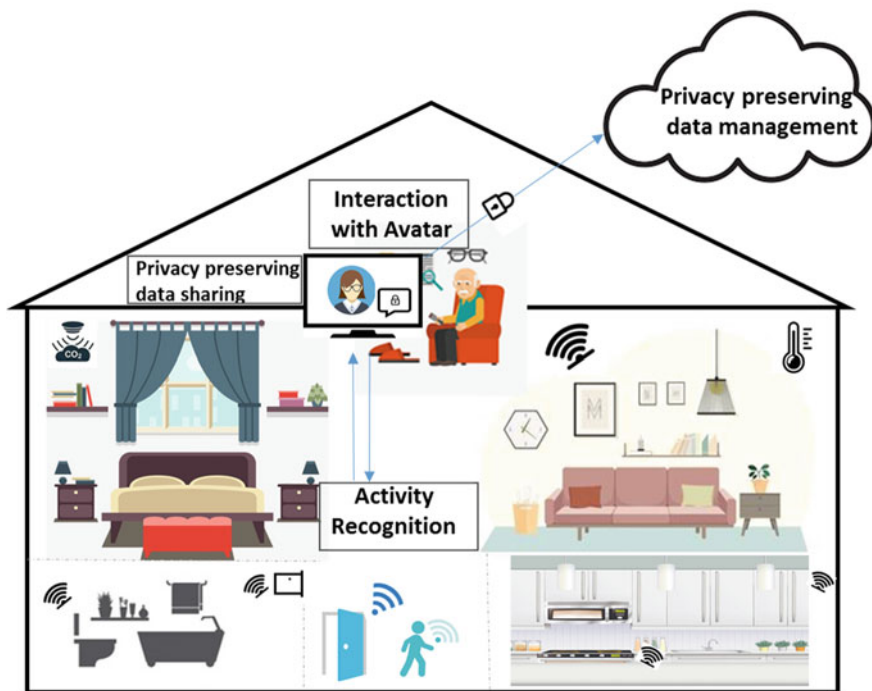


Fig. 16.1 Overview of the proposed framework

### ***16.3.1 Activity Recognition and Occupancy Detection Module***

The activity recognition module monitors daily living activities of the resident such as walking, sleeping, cooking, bathing and watching TV by using the unobtrusive sensor data from the home environment. The aim is to provide health care, comfort, safety and security to residents by detecting abnormal events and providing home automation services. The occupancy detection module of the framework identifies the resident's location and a number of occupants in home by using the data obtained from the sensors. The data obtained are then analysed and used to train machine learning algorithms for classification of various activities, resident behaviour analysis and detection of multiple occupancy in smart home.

One aspect often regarded is explainability in classification systems. Since many smart systems currently deploy deep learning models that are often treated as black box and can be hard to explain. Recent advances in explainability of deep learning systems will be also deployed in our proposed framework. Local Interpretable Model-Agnostic Explanations (LIME) [30] is widely used model-agnostic method to explain the model decision. It uses the trained model and by perturbing the input to the model it observes, how the prediction of model changes. LIME can be applied to image, categorical data or text. In our framework, since LIME is model-agnostic, we can deploy it for all classification tasks regardless of model specifications. Another method we will deploy is to explain our smart systems, which will be diving higher-level classification into smaller classification along with object and surrounding detection, which can give insight into why a certain decision is made. For example, if cooking is the main activity that is detected by our smart system, our system will also show low-level activities that are detected can be related to cooking. These low-level activities can be tap water is flowing, chopping sound detected, the oven is on, objects detected can be a knife, pan and potatoes and location of the user which can be the kitchen. These low-level activities can shed light, while cooking activity was detected also in the case or wrong classification of activity or not being able to classify an activity, again these low-level activities can explain the reason to the user. This methodology can be applied to all different part of the system whenever a decision is made by a smart algorithm.

### ***16.3.2 Privacy-Enabled Voice Assistant Module***

The privacy-enabled voice assistant allows the development of a more privacy-aware smart home that cannot only detect user activities and protect data. But it can also interact with the end-user in a meaningful way through a dialogue manager, which can learn and take into account their individual preferences not only about daily living activities but also about privacy. This includes interactions such as sometimes alerting users about data sharing they may not feel comfortable with, refining models

of their user's preferences over time and at times prompting the users to carefully consider the possible implications of some of their privacy decisions.

This module is able to learn the user's preferences and help them manage their smart home and privacy settings across a wide range of devices and sensors without the need for frequent interruptions. In addition, with the privacy-preserving data management module, the users can automatically anonymize the collected data at the desired granularity and distribute or upload to the cloud. Different subsets of the data that are required by a third-party application or according to the access level of the receiver can be shared without providing access to the completely original data set.

## 16.4 System Architecture

The section presents a detailed description of the integration of each module of the framework and how the interaction and communication between them are performed. In the proposed framework, the Home Event Recognition System (HOMER) [10] acts as a middleware to integrate each module and sensor components of the system. HOMER is an open-source platform based on the Apache Karaf OSGi framework and enables modularity by encapsulating its functionalities in terms of OSGi bundles. The framework is flexible, extendable and adaptable to new components/modules. The components are in the form of OSGi bundles that can be remotely installed and updated without rebooting the system.

Figure 16.2 shows the system architecture of the proposed framework. The detailed description of each component is explained below.

### 16.4.1 Sensor Data and Pre-processing

The input to the system is the raw sensor data of the user and the environment from the smart home. The sensors used are motion sensors, door, window sensors, light sensor, temperature, humidity, passive infrared (PIR), pressure mats and CO<sub>2</sub> sensors. The data obtained are represented by a sequence of time-value pairs such as  $\langle t_n, v_n \rangle, \langle t_{n+1}, v_{n+1} \rangle$  represents two consecutive pairs, where  $v_n$  spans the time interval  $[t_n, t_{n+1})$ . The sampling rate varies on per sample basis depending on the sensors. Therefore, it is important to perform data pre-processing as it has a significant impact on the performance of machine learning models.

The data pre-processing includes removal of noisy data and outliers using data filtering techniques and handling of missing/incomplete values, which can be performed using interpolation methods to maintain completeness and consistency in time series. In the smart home systems, it is also very important to check if the data received are from the right sensor of a wireless sensor network. After the data

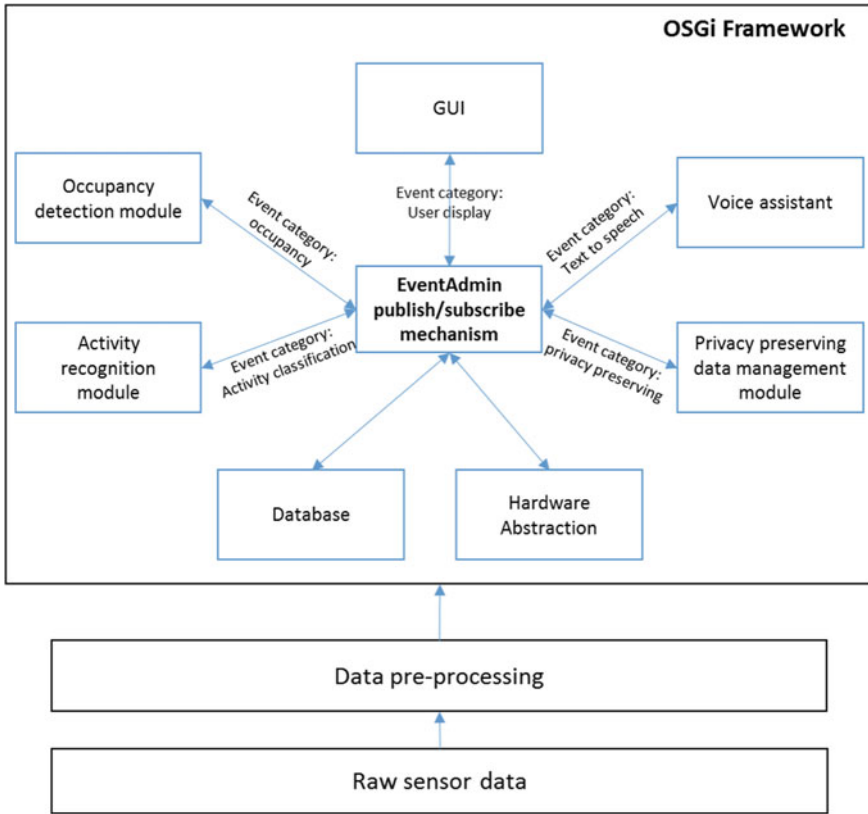


Fig. 16.2 System architecture

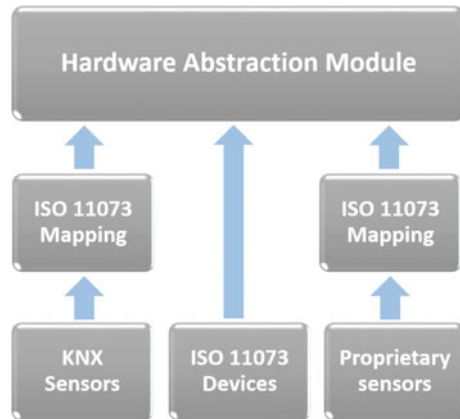
pre-processing is performed, the data are sent to the hardware abstraction bundle of OSGi framework.

The OSGi framework of HOMER ensures modularity and flexibility, which facilitates the parallel deployment of a bundle in the framework. Each bundle is separated and provides different functionalities. Each module of the OSGi framework is termed as bundles. As shown in Fig. 16.2, it consists of bundles, i.e. hardware abstraction, database, activity recognition, occupancy detection, data management, voice assistant and graphical user interface (GUI) bundles.

### 16.4.2 Hardware Abstraction

The hardware abstraction acts as an intermediate layer for harmonization of incoming sensor data from various sensor networks. It provides a mapping of non-standardized devices to the appropriate notation in ISO 11073 specification. The ISO/IEEE 11073

**Fig. 16.3** Layout of hardware abstraction module [10]



standards enable the system to exchange sensor data between different medical devices and systems analysing these data. The home automation sensors are covered in the Independent Living Activity Hub specialization ISO/IEEE 11073-10471 [10]. This module provides the possibility of integration of non-intrusive off-the-self devices from different domains for data acquisition. The layout of the hardware abstraction module consisting of different bundles is shown in Fig. 16.3.

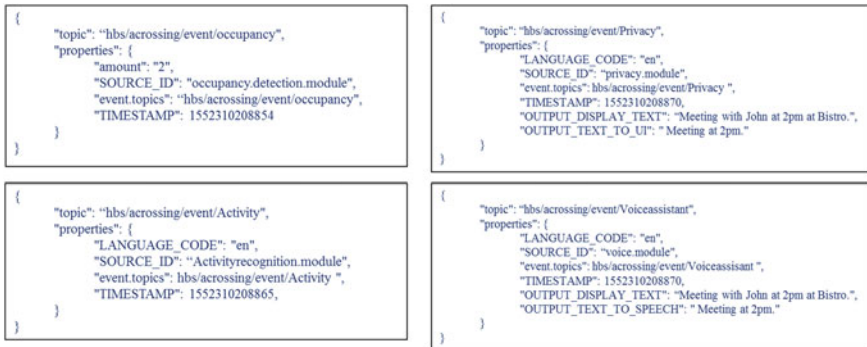
A hardware abstraction bundle is used as an abstraction layer between HOMER and different kinds of hardware. This enables hardware to distribute events generated by devices or to get commands for actuation. This way all types of hardware can use the same events for different devices.

### 16.4.3 Database

The database bundle stores all the sensor data in a standardized format. It contains information about the HouseID, RoomID, SensorID, sensor type, actuator ID, actuator and message type such as “ON” or “OFF”. A data access interface provides the functionality to query the database for information retrieval or persistence.

### 16.4.4 Event Admin

All the bundles of the framework are interconnected with the Event Admin. The Event Admin is supplied by the OSGi framework and provides communication between the bundles by sending and receiving asynchronous events. It is based on a subscribe/publish mechanism in which bundles need to subscribe to Event Admin in order to communicate and send/receive events within the framework. Each event is defined by a topic, which specifies the type of event and all the topics of the



**Fig. 16.4** Example of event admin service specification

framework use the same prefix for identification followed by a category identifier. An event can also consist of one or more properties. A property is consisted of a key and its corresponding value. Examples of events are shown in Fig. 16.4.

Each of the event topics is linked to a specific category. The event categories in the proposed framework are occupancy, activity, privacy and voice assistant. A category consists of various event topics together with a message and set of properties, which enables communication and interaction between the different bundles.

### 16.4.5 Activity Recognition

The activity recognition bundle of the framework gathers processed and clean data from the database and performs a classification of the different activities of the user. In order to train the classifier, we prefer deep learning models such as long short-term memory (LSTM) and convolutional neural network (CNN) as they have the capability to learn features automatically through the network instead of manual handcrafting, which is a time-consuming and costly process. The Event Admin receives the classification output from the bundle and sends it to other bundles whenever it is required.

### 16.4.6 Occupancy Detection

The occupancy detection bundle uses a fusion of multiple non-intrusive sensors which have shown to be higher in accuracy, detection and occupancy estimation [7]. We prefer CO<sub>2</sub> sensors, motion sensors, PIR, temperature, light, door/window sensors installed in a home environment for occupancy estimation.

The occupancy detection can be seen as a multi-class problem with the aim to estimate the number of occupants in a home, and in order to train classifiers, recurrent

neural network-based models have outperformed the existing standardized machine learning models. Since the data set can be unbalanced, we prefer F-score results to predict the performance of the classifiers. In case of multi-class problem, a macro-average method can be used, which is the average of precision and recall of each class label and then calculates the F-score.

### ***16.4.7 Privacy-Preserving Data Management***

The privacy-preserving data management bundle is used to sanitize the data before other modules access them. It anonymizes the data and removes personally identifiable information based on common personal information attributes as described in the Europe General Data Protection Regulation (GDPR). User preferences towards privacy have also taken into consideration, which we analysed according to the user studies [29].

In previous work [28], the LSTM encoder–decoder approach has been implemented where the encoder network takes the input sequence and maps it to an encoded representation (vector of fixed dimensionality). Then, the decoder network uses the encoded representation to generate output sequence. This makes the model a lock and key analogy, where only the correct key or decoder will be able to access the resources behind the lock (encoder). The results on simulated smart home data set showed that method is able to learn privacy operations such as disclosure, generalization and deletion of data; thus generating different data views for data receivers in an encrypted way.

### ***16.4.8 Voice Assistant***

The voice assistant bundle handles the interaction between the user and the smart home in an efficient and comfortable way. The system consists of sequences of processes in which input to the system user's utterance and in return produce spoken output. A general architecture of a spoken dialogue system described in [27] consists of various components: first, the automatic speech recognition (ASR) component converts the raw audio input into a sequence of words (or the n-best results), which is then forwarded to a natural language understanding (NLU) component to extract the semantics of the utterance. A formal representation of semantics, generally a structured set of attribute-value pairs is used by the dialogue manager (DM) to decide the next actions to take according to the employed dialogue strategy. The action performed by DM is the generation of text from user utterance and transformation of text into speech by the text-to-speech engine. Other actions of DM are interaction with the back-end system or any kind of processing required by the application.

In this proposed framework, the dialogue manager module handles the interaction between the user and the smart home. In previous work, dialogue managers were



developed using rule-based systems in order to be robust. The drawbacks of rule-based dialogue systems are that they require a human expert to design rules, which is time-consuming and makes the system static. Recently, a new model has been developed in a dialogue setting where dialogues are train with images instead of words [18]. This method handles words efficiently on which dialogue manager is not trained on and do not require pre-processing the data. We prefer this image based approach to train the dialogue manager.

### ***16.4.9 Graphical User Interface (GUI)***

The graphical user interface bundle provides graphical representation of the smart home architecture, which gives information about the user activities, statistics and analysis of activity data from the activity recognition module. It also provides information about user daily agenda/meetings and notifies the user for the upcoming event. The GUI of the system can be in the form of an avatar or mobile application on tablet, PCs or smartphone. All the information is presented in an abstract manner (for example: replacing the private activities such as bathing and sleeping with some unique notation; and anonymizing the personal information such as the name of the person with whom the meeting is scheduled, address and contact details) according to the privacy preferences of the user.

## **16.5 Example Use Cases of the Proposed Framework**

The following section presents two different scenarios to illustrate how all the modules of the framework are interacting with each other in order to provide assistance to the resident.

### ***16.5.1 Scenario 1***

The first scenario examines the case where there is a reminder for a meeting that needs to be communicated to the user. In homes with multiple residents, it is a common phenomenon that there are different privacy preferences or events that people do not want to be shared. Therefore, in this instance, the framework makes use of the occupancy detection module to detect if the user is alone before communicating the details of the upcoming meeting. As a first step, the sensors installed in the home environment collect the data over a continuous period. Specifically, the sensors that were selected give data about the resident's location, the number of residents (one or more than one) and daily activities performed by the residents. Then, the dialogue manager module and privacy management module select the appropriate information

**Table 16.1** Meeting reminder

|                      |   |
|----------------------|---|
| Description          | The use case describes the steps associated with the meeting reminder service. The voice assistant reminds the user of an upcoming meeting and depending on the circumstances reveals only specific information   |
| Actors               | User and voice assistant  |
| Pre-conditions       | The voice assistant has knowledge of the occupancy status and the events in the calendar of the user  |
| Post-conditions      | The user is reminded of an upcoming appointment without excess information being given in the presence of other people  |
| Action sequence      | <ol style="list-style-type: none"> <li>1. The service starts with the framework detecting the occupancy status</li> <li>2. If more than one person is present in the room, the voice assistant enables the privacy module, which hides not essential information</li> <li>3. The voice assistant reminds the user of the upcoming meeting but does not mention the location, the identity of the other participants and the purpose of the meeting</li> </ol> |
| Alternative sequence | <ol style="list-style-type: none"> <li>1. The framework detects the occupancy status and the user is alone</li> <li>2. The voice assistant reveals all the related information for the upcoming meeting</li> </ol>  |
| Requirements         | The voice assistant has access to the smart home sensor data and the user's agenda  |

that needs to be communicated based on the result of the occupancy detection. Lastly, the voice assistant uses the dialogue manager to communicate the reminder to the user. Table 16.1 shows the procedure and interaction followed by the voice assistant.

## 16.5.2 Scenario 2

The second scenario examines the case of having guests in the smart home. There are many cases in which guests are not comfortable if they feel they are being monitored and they would like to be informed about the type of monitoring that takes place. User preferences show the users' expectations in regard to how their data should be managed by the smart home. The voice assistant can meet these preferences partially or completely depending on service policies and other user preferences existing in the same space. For example, user preferences in this scenario are:

- Preference 1: Do not share the occupancy status of my house when guests are present.  
 Preference 2: Do not share my location with anyone.

The framework is able to deal with these privacy preferences by first detecting the presence through the occupancy detection module. The dialogue manager prompt user informs multiple occupancy and inquires user about their preferences towards monitoring. If the user disagrees with monitoring, then the privacy-preserving data management module applies user preferences in the privacy policy of the module and

**Table 16.2** Guests privacy

|                      |  |
|----------------------|--|
| Description          | The use case describes the steps associated with the guest service. The voice assistant adjusts his privacy module to allow for the guests' privacy preferences  |
| Actors               | User, guests and voice assistant   |
| Pre-conditions       | The voice assistant has knowledge of the occupancy status  |
| Post-conditions      | The guests are informed about the presence of sensors and their privacy preferences are taken into account for the data collection and processing  |
| Action sequence      | <ol style="list-style-type: none"> <li>1. The framework detecting the occupancy status and the presence of guests in the smart home</li> <li>2. The voice assistant makes them aware of the sensors and that data collection occurring in the smart home and asks them if they agree with the monitoring</li> <li>3. The guests do not feel comfortable with monitoring and they would like to have their data removed</li> <li>4. The framework annotates the occupancy data during the time the guests were in the smart home and they are obfuscated before being processed by third parties</li> <li>5. The voice assistant informs the guests that their data will be removed and turns itself off (not listening mode) until the occupancy module detects that the guests have left</li> </ol> |
| Alternative sequence | <ol style="list-style-type: none"> <li>1. The guests reply that they do not mind the monitoring as long as their identity remains hidden.</li> <li>2. The framework annotates the data that was collected during that time as data that should be anonymized before shared with third parties</li> </ol>   |
| Requirements         | Voice assistant has access to the smart home data of the resident and multiple occupancy   |

removes or anonymizes the data for some time until the user agrees with monitoring again. Table 16.2 presents the procedure and implementation of scenario 2. In both cases, the privacy-enabled voice assistant is able to handle the different privacy preferences of the users through its dialogue manager and the private data management module. This is a feature that has become more important lately and should be incorporated in all frameworks, as people become more aware of privacy and potential risks to it they have higher privacy expectations for the systems they use.

## 16.6 Conclusion and Prospects

The tremendous growth in smart home technology has led to improved home care through daily monitoring and automation. The data collected in the home environment raise various privacy and security concerns among the users. Therefore, in this chapter, we presented a secure smart home, which not only detects daily living activities of the user but also protects data and performs an interaction with the user in a meaningful way through the dialogue manager. The chapter presents the

major components of the framework: activity recognition and occupancy detection; privacy-preserving data management and dialogue manager and propose methodologies to develop these modules. A system architecture of the proposed framework describes the communication and interaction between the modules in the HOMER middleware. One of the innovative aspects of the proposed framework is privacy-enabled voice assistant, which learns and takes into account users' preferences while sharing information and anonymizes the sensitive and personal information of the user depending upon the situation. In future work, implementation of the proposed framework will be performed and major focus will be on the integration of the individual modules in the OSGi framework so that system will be scalable, flexible and adaptable according to the different applications.

**Acknowledgements** This work has been funded by the European Union Horizon2020 MSCA ITN ACROSSING project (GA no. 616757). The authors would like to thank the members of the project's consortium for their valuable inputs.

## References

1. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6:52138–52160
2. Ahmad J, Larijani H, Emmanuel R, Mannion M, Javed A (2018) Occupancy detection in non-residential buildings—a survey and novel privacy preserved occupancy monitoring solution. *Appl Comput Inform*. <https://doi.org/10.1016/j.aci.2018.12.001>
3. Apthorpe N, Huang DY, Reisman D, Narayanan A, Feamster N (2018) Keeping the smart home private with smart(er) IoT traffic shaping. arXiv preprint arXiv:181200955
4. Bos JW, Lauter K, Naehrig M (2014) Private predictive analysis on encrypted medical data. *J Biomed Inform* 50:234–243
5. Chen H, Liu X, Yin D, Tang J (2017) A survey on dialogue systems: recent advances and new frontiers. *ACM SIGKDD Explor Newsl* 19(2):25–35
6. Chen L, Nugent C, Okeyo G (2014) An ontology-based hybrid approach to activity modeling for smart homes. *IEEE Trans Hum Mach Syst* 44(1):92–105
7. Chen Z, Jiang C, Xie L (2018) Building occupancy estimation and detection: a review. *Energy Build* 169:260–270
8. Das A, Degeling M, Wang X, Wang J, Sadeh N, Satyanarayanan M (2017) Assisting users in a world full of cameras: a privacy-aware infrastructure for computer vision applications. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), IEEE, pp 1387–1396
9. Dwork C (2011) Differential privacy. In: *Encyclopedia of cryptography and security*, pp 338–340
10. Fuxreiter T, Mayer C, Hanke S, Gira M, Sili M, Kropf J (2010) A modular platform for event recognition in smart homes. In: The 12th IEEE international conference on e-health networking, applications and services. IEEE, pp 1–6
11. GDPR (2018) General Data Protection Regulation (GDPR) final text neatly arranged. [online] Available at: <https://www.gdpr-info.eu/>. Accessed 16 Apr 2019
12. Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:171209923
13. Holzinger A, Kieseberg P, Weippl E, Tjoa AM (2018) Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI.

- In: International cross-domain conference for machine learning and knowledge extraction. Springer, Cham, pp 1–8
14. Jia R, Dong R, Sastry SS, Sapanos CJ (2017) Privacy-enhanced architecture for occupancy-based HVAC control. In: 2017 ACM/IEEE 8th international conference on cyber-physical systems (ICCPS), IEEE, pp 177–186
  15. Jung Y (2017) Hybrid-aware model for senior wellness service in smart home. *Sensors* 17(5):1182
  16. Liu B, Andersen MS, Schaub F, Almuhimedi H, Zhang SA, Sadeh N, Agarwal Y, Acquisti A (2016) Follow my recommendations: a personalized privacy assistant for mobile app permissions. In: Twelfth symposium on usable privacy and security (SOUPS 2016), pp 27–41
  17. Machado E, Singh D, Cruciani F, Chen L, Hanke S, Salvago F, Kropf J, Holzinger A (2018) A conceptual framework for adaptive user interfaces for older adults. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), IEEE, pp 782–787
  18. Merdivan E, Vafeiadis A, Kalatzis D, Henke S, Kropf J, Votis K, Giakoumis D, Tzovaras D, Chen, L, Hamzaoui R, Geist M (2018) Image-based natural language understanding using 2D convolutional neural networks. arXiv preprint arXiv:181010401
  19. Mittelstadt B, Russell C, Wachter S (2018) Explaining explanations in AI. arXiv preprint arXiv:181101439
  20. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):p529
  21. Monteriù A, Prist M, Frontoni E, Longhi S, Pietroni F, Casaccia S, Scalise L, Cenci A, Romeo L, Berta R, Pescosolido L (2018) A smart sensing architecture for domestic monitoring: methodological approach and experimental validation. *Sensors* 18(7):p2310
  22. Naehrig M, Lauter K, Vaikuntanathan V (2011) Can homomorphic encryption be practical? In: Proceedings of the 3rd ACM workshop on cloud computing security workshop, ACM, pp 113–124
  23. Okeyo G, Chen L, Wang H (2014) Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes. *Future Gener Comput Syst* 39:29–43
  24. Park Y, Kang S, Seo J (2018) An efficient framework for development of task-oriented dialog systems in a smart home environment. *Sensors* 18(5):1581
  25. Pathak M, Rane S, Sun W, Raj B (2011) Privacy preserving probabilistic inference with hidden Markov models. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5868–5871
  26. Pathak MA, Raj B (2013) Privacy-preserving speaker verification and identification using gaussian mixture models. *IEEE Trans Audio Speech Lang Process* 21(2):397–406
  27. Pieraccini R, Huerta J (2005) Where do we go from here? Research and commercial spoken dialog systems. In: 6th SIGdial workshop on discourse and dialogue
  28. Psychoula I, Merdivan E, Singh D, Chen L, Chen F, Hanke S, Kropf J, Holzinger A, Geist M (2018) A deep learning approach for privacy preservation in assisted living. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom workshops), IEEE, pp 710–715
  29. Psychoula I, Singh D, Chen L, Chen F, Holzinger A, Ning H (2018) Users' privacy concerns in IoT based applications. In: 2018 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI), IEEE, pp 1887–1894
  30. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1135–1144
  31. Serban IV, Lowe R, Henderson P, Charlin L, Pineau J (2015) A survey of available corpora for building data-driven dialogue systems. arXiv preprint arXiv:151205742

32. Singh D, Merdivan E, Hanke S, Kropf J, Geist M, Holzinger A (2017) Convolutional and recurrent neural networks for activity recognition in smart environment. In: *Towards integrative machine learning and knowledge extraction*. Springer, Cham, pp 194–205
33. Singh D, Merdivan E, Psychoula I, Kropf J, Hanke S, Geist M, Holzinger A (2017) Human activity recognition using recurrent neural networks. In: *International cross-domain conference for machine learning and knowledge extraction*. Springer, Cham, pp 267–274
34. Singh D, Psychoula I, Kropf J, Hanke S, Holzinger A (2018) Users' perceptions and attitudes towards smart home technologies. In: *International conference on smart homes and health telematics*. Springer, Cham, pp 203–214
35. Sivaraman V, Gharakheili HH, Vishwanath A, Boreli R, Mehani O (2015) Network-level security and privacy control for smart-home IoT devices. In: *2015 IEEE 11th international conference on wireless and mobile computing, networking and communications (WiMob)*, IEEE, pp 163–167
36. Sutskever I, Vinyals O, Le, QV (2014) Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*, pp 3104–3112
37. Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driven user-centric explainable AI. In: *Proceedings of the SIGCHI conference on human factors in computing systems CHI*, vol 19
38. Wang J, Chen Y, Hao S, Peng X, Hu L (2019) Deep learning for sensor-based activity recognition: a survey. *Pattern Recogn Lett* 119:3–11
39. Xie P, Bilenko M, Finley T, Gilad-Bachrach R, Lauter K, Naehrig M (2014) Crypto-nets: neural networks over encrypted data. *arXiv preprint arXiv:14126181*
40. Yang J, Zou H, Jiang H, Xie L (2018) Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes. *IEEE Internet Things J* 5(5):3991–4002
41. Zheng S, Apthorpe N, Chetty M, Feamster N (2018) User perceptions of smart home IoT privacy. *Proc ACM Hum-Comput Interact* 2(CSCW):200

# Index

## A

- Accelerometry, 172
- Active assisted living, 40, 91, 119, 209
- Activities of daily living, 23, 56, 120, 121, 126, 152, 255
- Activity classification, 37, 119, 122, 124–126, 139, 142
- Activity recognition, 3, 7–9, 11–13, 15, 16, 115, 120, 122, 124, 125, 133, 148, 150–152, 154, 156, 169, 172, 228–230, 232–235, 238, 243, 244, 250–254, 262, 264, 322–324, 328, 330, 332, 334, 337
- Affect recognition, 211, 213, 224
- Ambient assisted living, 120, 270
- Ambient intelligence, 150, 250, 251, 278, 281, 306
- Application mark-up, 239
- Assistive system set-up, 227–230, 233, 234, 237–239, 243, 244

## B

- Batch normalization, 93, 108–110, 115
- Behavioral assistance, 228, 233

## C

- Context model, 228–234, 238, 242–244
- Context sensing, 13
- Convolutional Neural Network (CNN), 14, 15, 91, 93, 96, 109–116, 123–125, 129, 131, 332

© Springer Nature Switzerland AG 2020

F. Chen et al. (eds.), *Smart Assisted Living*, Computer Communications and Networks, <https://doi.org/10.1007/978-3-030-25590-9>

## D

- Data entry, 63, 64, 66–68, 70–76, 78, 79, 81, 82, 84
- Data reduction, 311, 313, 316, 318
- Dataset quality, 11, 13, 39, 92, 93, 108–110, 113, 114, 121–129, 132, 135, 137, 139, 142, 147–150, 152, 155–158, 160–163, 170, 174, 175, 180, 184, 291, 292, 294, 296, 297, 300, 305, 306, 308, 311–317, 325, 333
- Deep learning, 11, 14, 55, 92, 93, 95, 96, 108, 110, 114, 169, 181, 251, 252, 254, 323, 325, 328, 332
- Detection quality, 124, 135
- Dialogue manager, 326, 328, 333–337

## E

- Egocentric vision, 120, 123
- Eyewear computer, 63, 64, 66, 68, 83

## F

- Face detection, 313, 314, 316, 317
- Features selection, 154

## G

- Gait, 39, 44, 49, 150
- Galvanic Skin Response (GSR), 213, 305, 307, 308, 311, 313–318
- Geo-data mining, 308, 313, 317

## H

- Health, 8, 13, 17, 44, 51, 56, 57, 64, 67, 82, 148–150, 169, 172, 185,

- 191, 194–197, 199, 201, 203, 205, 236, 249, 250, 252–256, 269–274, 278–280, 283, 284, 289, 302, 306, 307, 328
- Human Activity Recognition (HAR), 8, 10, 11, 91, 92, 94, 95, 109, 113, 148–150, 152, 153, 155, 162, 163, 169, 172, 228, 230–232, 238, 243, 323
- Human activity tracking, 23
- I**
- Indoor localization, 3, 4, 6, 11, 23, 24, 26, 123
- Information and Communication Technology (ICT), 148–150, 191–196, 199, 201, 204, 206, 234
- Innovation, 67, 143, 185, 204, 206, 252, 270–272, 274–278, 282–284, 302
- L**
- Life-log, 306–317
- Living labs, 269–283
- Location classification, 122–124, 126, 133, 136, 141, 142
- M**
- Machine learning, 14, 18, 39, 93, 116, 120, 124, 125, 149, 155, 156, 162, 169, 170, 172, 175, 177, 178, 181, 183, 184, 257, 322, 324, 325, 328, 329, 333
- MapReduce, 288–290, 292, 294, 296–299
- Market access, 200, 275, 280
- Microsoft Kinect, 25
- Mild cognitive impairment, 23
- Multi-resident activity, 249, 251, 253, 254
- O**
- Object detection, 24, 55, 119, 124–126, 129, 132, 133, 135–139, 141, 142
- Occupancy detection, 323, 324, 328, 330, 332, 334, 335, 337
- P**
- Perinatal stroke, 167, 168, 170, 173, 183, 184
- Pervasive sensing, 3, 19, 169
- Physiological, 17, 18, 307, 308, 311, 314, 316, 317
- Post-stroke, 43, 44, 47, 48, 50, 51, 57, 60
- Prechtl's general movements assessment, 168, 171
- Privacy preserving data management, 323, 329, 333, 335, 337
- R**
- Radar, 18, 44, 51–53, 55–60
- Rehabilitation, 25, 44, 48–51, 56, 57, 60, 120, 192, 193, 201, 210, 280
- Robotic assistant, 224
- S**
- Self-report, 84, 305, 311–314, 316, 317
- Semantic manual, 244
- Sensors, 3, 7, 8, 10–12, 14, 15, 18, 23–27, 29–37, 39, 40, 43, 44, 51–60, 63, 64, 68, 92, 120, 123, 147, 149–154, 212, 163, 167, 169, 170, 172–174, 184, 212, 213, 228, 230, 233–235, 238, 243, 244, 249–257, 260, 261, 305–308, 310, 314, 317, 321–324, 326–332, 334–336
- Smart home, 19, 63, 64, 91, 116, 148, 234, 235, 249–254, 256–258, 262, 264, 287, 322–325, 327–329, 333–336
- Smart medical, 288–290, 295, 301, 302
- Spatial kNN, 289, 301
- T**
- Temporal associations, 125, 141
- Thermal, 44, 51, 53, 56–60
- Transfer Learning, 8, 9, 96, 98–100, 102
- U**
- UCF101, 92, 93, 105, 108–110, 113–116
- Unobtrusive, 43, 51–54, 56, 57, 60, 253, 309, 323, 328
- User-centered design, 191, 192
- V**
- Validation and evaluation, 264, 275
- VGG16, 97, 99, 100, 103, 109, 110, 113–116
- Voice-based annotation, 254, 257
- W**
- Wearable, 8, 11, 14, 26, 27, 44, 51, 56, 57, 60, 63, 64, 67, 68, 73, 120, 123, 150, 151, 167, 169, 170, 172, 196, 228, 230, 235, 250, 255, 288, 306–309, 323
- Well-being, 23, 191, 195, 210, 250, 254, 278, 281, 283