Jihad Mohamad Alja'am
Abdulmotaleb El Saddik
Abdul Hamid Sadka *Editors*

Recent Trends in Computer Applications

Best Studies from the 2017 International Conference on Computer and Applications, Dubai, UAE

Springer

Recent Trends in Computer Applications

Jihad Mohamad Alja'am •
Abdulmotaleb El Saddik • Abdul Hamid Sadka
Editors

Recent Trends in Computer Applications

Best Studies from the 2017 International Conference on Computer and Applications, Dubai, UAE

Springer

Editors

Jihad Mohamad Alja'am
Computer Science and Engineering
Qatar University
Doha, Qatar

Abdulmotaleb El Saddik
Faculty of Engineering
University of Ottawa
Ottawa
Ontario, Canada

Abdul Hamid Sadka
Electronic and Computer Engineering
Brunel University London
Uxbridge, United Kingdom

ISBN 978-3-319-89913-8 ISBN 978-3-319-89914-5 (eBook)
https://doi.org/10.1007/978-3-319-89914-5

Library of Congress Control Number: 2018958782

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

Foreword

By the turn of the second millennium, it became clear that computers (and more broadly intelligent machines) are becoming the focus of science and technology for the next few decades to come. This book introduces the reader to the realm of the most recent trends in the area of computer applications, with a special focus on sustainable development, marking this important trend during the first decades of the third millennium. The broad scope of the book is by design as the editors and authors introduce a wide scope of application fields where modern computing brought about several paradigm shifts in the way data is analysed, managed and visualised.

Already a decade ago, the notion of Big Data was introduced, and since then new scientific and technical challenges were formulated and efficient solutions have been proposed. Big Data was later cast in the framework of decision-making environments, in which theories, algorithms, methods and systems have been developed to efficiently map data into decisions. Data-driven and data-intensive computer applications have since been developed in a number of areas, including, but not limited to, media (both audio and visual), healthcare, robotics, security, web applications and web interfaces. Conceptually, data handling strategy can conveniently be presented as a three-layered scheme, in which the first layer interfaces with raw data (computer-generated, time series, sensor data, etc.) and offers various ways to represent, clean, abstract and possibly augment the source data. The second layer hosts methods and algorithms for analytics, management and visualisation of the processed data. And finally, the third layer links the results of the second layer to a specific application, that is, it interfaces with the real-world application domain. Most of the contributions in this book cover one or more of these layers targeting a specific application domain. As the amount of data keeps increasing exponentially and the demand of split (real-time) decision is becoming more imminent, the key challenges we are facing today include scalability, efficiency and real-time performance.

The book is recommended to the tech-savvy managers as well as engineers, technicians and researchers in various fields of computer applications. It is rather seldom we come across a reference where such an overwhelming amount of information regarding diverse application fields has been gathered in a single volume. The book requires common mathematical and computer science knowledge acquired by most university college degrees and, therefore, it is easy to read and grasp the main concepts which are well illustrated throughout the chapters of the book.

Moncef Gabbouj

NSF I/UCRC Center for Visual
and Decision Informatics, TUT-Site
Laboratory of Signal Processing
Tampere University of Technology Tampere, Finland

Preface

This book consists of an agglomeration of know-how and recent research findings imparted by a broad range of international scientists within the field of information and communication technologies. The book recognises computers potentially as data-generating machines and computer applications as platforms for the acquisition, analysis, processing, management and visualisation of data in its multiple forms, scales (i.e. volumes), complexities and digital representations. While the editors realise the breadth and diversity of computer applications, and hence the corresponding data-handling strategies employed therein, the essence of the book focuses mainly on three key data-driven classes of technologies, namely data analytics, data management and data visualisation.

In data analytics and processing, the book presents an authoritative set of chapters addressing various challenges commonly encountered in computer-based applications and systems, such as segmentation, detection, classification, recognition, etc., in the light of vision-based but also multimodal scenarios. In image segmentation, for instance, one chapter addresses the unsupervised segmentation of images using graph-based community detection. In particular, an overview of sequential mining algorithms and their extensions is presented in one chapter, while image/video classification is addressed using multimodal techniques in another chapter and using Gabor filters in yet another. An analysis of the current and future directions of object detection based on convolutional neural networks is also featured. Hand detection and gesture recognition within the context of human–computer interaction is addressed in a chapter that focuses on translating recognised hand gestures into functional ones to enable the real-time manipulation of a 2D image. The book also looks into the compression aspects of computer applications with multiview video codecs in perspective. In particular, one chapter addresses the multiview extensions of two contemporary video coding standards and provides a comparative analysis of their performance in terms of quality and compression efficiency.

In data management, the authors' contributions place a particular emphasis on the security aspects of data in networked computer applications which utilise cloud computing technologies. One chapter looks at utilising a combination of data encryption algorithms and a distributed system to improve data confidentiality for an acceptable overhead performance. Another chapter considers the malware detection algorithms and argues the complexity and computationally intensive nature of the process of identifying malicious codes in files or network traffic. It goes further to propose a novel hybrid solution that leverages the CPU/GPU computing capabilities for improving the performance and reducing the power consumption of string matching algorithms on devices such as laptops for instance. Furthermore, the security aspects of Software-Defined Networking (SDN) are examined from the standpoint of Distributed Denial of Service (DDoS) attacks. The chapter presents a controller placement model that helps reduce the impact of DDoS attacks and hence make SDN more secure and resilient. One interesting computer application considered in the book is the control of a permanent-magnet DC motor without the prior knowledge of its parameters. Another chapter dedicates specific attention to designing a new hardware/software platform that enables the real-time provision of all the parameters required for the control of the DC motor.

In data access and visualisation, the book presents a series of chapters that are concerned with providing a user-centric approach to multimedia applications and services, particularly in e-commerce, future Tactile Internet, search and retrieval as well as language recognition scenarios. One chapter reviews the state of the art in user-centric multimodal systems and presents a vision towards the realisation of an immersive, interactive and collaborative framework for the Internet of Multimodal Things (IoMT) system. Another chapter considers an automated approach to the optimisation of Web interfaces for e-commerce. The chapter emphasises primarily the vital role of User Experience (UX) and Customer Experience (CX) principles in the provision of any web-based service, application or product. The book embodies a chapter that features the design, development and evaluation of a web-based Arabic multimedia search engine that is based on a language transcriber. In order to enable an efficient and user-friendly human–computer interaction, a chapter focuses on the review and analysis of specific text-to-picture systems and approaches to facilitate education. Last but not least, one of the chapters explores both person-dependent and person-independent Arabic speech recognition systems and examines how hidden Markov models can be specifically exploited for the recognition of Arabic, rather than English, words.

This book offers the readers with the unique dual benefit of gaining a meticulous analysis of current technology trends in computer applications and simultaneously benefiting from a rich display of recent experimental research findings in a rather diverse and prolific technological field. While the book is inherently diverse in its scope and coverage, addressing a broad spectrum of technologies exploited by computer applications and systems today, the book editors are confident that this manuscript will put at the disposal of their audience, from both academia and industrial R&D sectors, a useful resource that will not only help expand the beneficiaries' knowledge base in the relevant fields but will also offer them a supportive guide that is equipped with a sufficient level of scientific originality, depth and rigour into a cluster of technological trends and most recent research developments in multimedia data handling and manipulation, with computer applications in perspective.

Jihad Mohamad Alja'am
Abdulmotaleb El Saddik
Abdul Hamid Sadka

Acknowledgements

This book would not have been possible without the hard work of all the authors. I am especially indebted to Prof. Abdul Hamid Sadka and Prof. Abdulmotaleb El Saddik who worked hard to select the best submitted chapters to be included in this book. I am grateful to all reviewers who spent a lot of time in the peer-review process to improve the quality of the accepted chapters. Each reviewer has provided excellent and professional scientific guidance for the authors to refine their works to meet the scientific standard. I am indebted to Dr. Nabil Khelifi (Senior Editor—Springer) who trusted me to lead this project successfully. Many thanks also for Reyhaneh Majidi (Springer DE) who has always been very supportive. Finally, I really appreciate the hard work of my assistant Dana Bandock who allocated an important part of her time in the management for the successful completion of this book.

Department of Computer Science and Engineering Jihad Mohamad Alja’am
College of Engineering
Qatar University
Doha, Qatar

Contents

Part I Data Analytics and Processing (Including Classification, Compression, Segmentation, Mining, Detection and Recognition etc.)

Part II Data Handling and Management (Including Data Security, Database Handling, Cloud Computing, Hardware and Software Technologies)

Part III Data Access and Visualisation: Tools and Platforms/Systems

Part I
Data Analytics and Processing (Including Classification, Compression, Segmentation, Mining, Detection and Recognition etc.)

Overview on Sequential Mining Algorithms and Their Extensions

Carine Bou Rjeily, Georges Badr, Amir Hajjam Al Hassani, and Emmanuel Andres

1 Introduction

Interesting sequential patterns (SPs) in a sequence database are extracted using Sequential Pattern Mining algorithms. These patterns help in analyzing data and obtaining interesting and valuable knowledge from large amounts of data. Other techniques including Sequence Prediction and Sequential Rule Mining are also used nowadays for decision-making purposes. The main idea is to extract frequent subsequences, called patterns, from a massive amount of collected data and understand the relation(s) between these patterns. Many sectors are interested in these techniques. For example, analyzing customers' purchases to improve marketing strategy: Let's say a customer buys a camera and a lens. The next time he comes, he buys a tripod. That information could be used to predict customers' needs by understanding their interests. The company may then offer a tripod or a discount when buying a camera and a lens. Nowadays, Sequential Pattern Mining algorithms play an important role in the medical domain, for the notion of time is important in analyzing data related to patients or hospitals. Novel applications are based on sequential mining for decision-making in the medical field, such as in [1–4]. Sequence Prediction was also used to predict heart failure in [5, 6].

C. Bou Rjeily · A. H. Al Hassani
Nanomedicine Lab, Université de Bourgogne Franche – Comté, Belfort, France
e-mail: carine.bourjeily@utbm.fr; amir.hajjam-el-hassani@utbm.fr

G. Badr (✉)
TICKET Lab, Antonine University, Baabda, Lebanon
e-mail: georges.badr@ua.edu.lb

E. Andres
Université de Strasbourg, Centre Hospitalier Universitaire, Strasbourg, France
e-mail: emmanuel.andres@chru-strasbourg.fr

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_1

The first part of this chapter defines important terms and notations in the field. The second shows a survey on the most important and recent sequential mining algorithms according to a clear classification. Lastly, the chapter concludes with a classification tree showing the main categories of the algorithms and their extensions. It is important to know that this chapter provides the essential definitions and functionalities of the algorithms. Knowing the appropriate outputs of the algorithms will help the user in choosing the most efficient one for his/her studies.

2 Important Terms and Notations

Before presenting the algorithms and their classification, it is important to define some basic terms used in Sequential Pattern Mining in order to understand the mining process. These terms are commonly used in data mining processes and especially in Sequential Pattern Mining.

1. An item is an entity that can have multiple attributes: date, size, color, and so on.
2. $I = \{i_1, \ldots, i_n\}$ is a nonempty set of items. A *k*-itemset is an itemset with *k* items.
3. A sequence "*S*" is an ordered list of itemsets. An itemset X_y in a sequence, with $1 \leq y \leq L$, is called a transaction. L denotes the length of the sequence, which refers to the number of its transactions. $S = \{(a,b); (b,c); (e,d)\}$, which means that the items a and b are occurring together in the same time, while the items b and c are occurring together although in the same time but after a and b occur together and so on.
4. A sequential database (SDB) is a list of sequences with a sequence ID (SID) (cf. Table 1).
5. A sequence β can have a subsequence α, making β a super-sequence of α.
6. A sequential rule r, denoted $X \rightarrow Y$, is a relationship between two unordered itemsets $X, Y \subseteq I$, where $X \cap Y = \varnothing$. $X \rightarrow Y$ means that if items of X appear in a sequence, items of Y will also occur in the same sequence.
7. The support of a rule r in a sequence database SDB is defined as the number of sequences that contains $X \cup Y$ divided by the number of sequences in the database:

$$supSDB(r) = \frac{|\{s; s \in SDB \wedge r \wedge s\}|}{|SDB|}$$

Table 1 A sequence database

SID	Sequence
1	$\langle \{a, b\}, \{c\}, \{f, g\}, \{g\}, \{e\} \rangle$
2	$\langle \{a, d\}, \{c\}, \{b\}, \{a, b, e, f\} \rangle$
3	$\langle \{a\}, \{b\}, \{f\}, \{e\} \rangle$
4	$\langle \{b\}, \{f, g\} \rangle$

8. The confidence of a rule r in a sequence database SDB is defined as the number of sequences that contains , divided by the number of sequences that contains X:

$$conf\,SDB(r) = \frac{|\{s; s \in SDB \wedge r \vee s\}|}{|SDB|}$$

9. A rule r is a frequent sequential rule iff $supSDB(r) \geq minsup$, with $minsup \in [0, 1]$ being a threshold set by the user.
10. A rule r is a valid sequential rule iff it is frequent and $confSDB(r) \geq minconf$, with $minconf \in [0, 1]$ being a threshold set by the user.
11. Apriori-based [7]: many mining algorithms are based on this technique. The main idea is to create a list of the most frequent items with respect to *minsup* and *minconf*. The list is increased progressively considering the support and the confidence.
12. Sequential Rule Mining is to find all frequent and valid sequential rules in an SDB [8].
13. Pattern Growth [9] is a method for extracting frequent sequences by partitioning the search space and then saving the frequent itemsets using a tree structure. Extraction is done by concatenating to the processed sequence (called prefix sequence) frequent items with respect to its prefix sequence. This method can be seen as depth-first traversal algorithm and eliminates the necessity to repetitively scan all of the SDB.
14. Searching processes:

    - Depth-First Search (DFS) is a searching process that traverses or searches tree or graph data structures. A node in the graph or tree is considered as the root where the search begins. In case of graph, some arbitrary nodes are selected as the root and explored as far as possible along each branch before backtracking.
    - Breadth-First Search (BFS) is a searching process for searching in trees or graph structures. It starts at the root like (DFS) and explores the neighbor nodes first, before exploring the next-level neighbors.

    Let $\beta = \langle\, \beta_1\ \ldots\ \beta_\mathrm{n} \,\rangle$ and $\alpha = \langle\, \alpha_1\ \ldots\ \alpha_\mathrm{m} \,\rangle$ be two sequences where $m \leq n$.
15. Sequence α is called the prefix of β iff ∀i ∈ $[1 \ldots m]$, $\alpha_i = \beta_i$.
16. Sequence $\beta = \langle\, \beta_1\ \ldots\ \beta_\mathrm{n} \,\rangle$ is called the projection of some sequence S with regards to α, iif:

    - $\beta \preceq s$
    - α is a prefix of β
    - There exists no proper super-sequence β' of β such that $\beta' \preceq s$ and β' also has a prefix

17. Sequence $\gamma = \langle\, \beta_{m+1}\ \ldots\ \beta_n \,\rangle$ is called the suffix of s with regard to α. β is then the concatenation of α and γ.

    Let SDB be a sequence database.

Table 2 A vertical database for the sequence database of Table 1

A		B		C		D		E	
SID	Itemsets	SID	Itemsets	SID	Itemsets	SID	Itemsets	SID	Itemsets
1	1	1	1	1	2	1		1	5
2	1,4	2	3,4	2	2	2	1	2	4
3	1	3	2	3		3		3	4
4		4	1	4		4		4	
		F				G			
		SID	Itemsets			SID	Itemsets		
		1	3			1	3,4		
		2	4			2			
		3	3			3			
		4	2			4	2		

Table 3 Projected database with regards to prefix "*a*"

$\langle a \rangle$: projected database
$\langle$ {_, b}, {c}, {f, g}, {g}, {e} $\rangle$
$\langle$ {_, d}, {c}, {b}, {a, b, e, f} $\rangle$
$\langle$ {b}, {f}, {e} $\rangle$
$\langle\ \rangle$

18. Horizontal database: each entry in a horizontal database is a sequence as shown in Table 1.
19. Vertical database: each entry represents an item and indicates the list of sequences where the item appears and the position(s) where it appears [10] (cf. Table 2).
20. Projected database: the α-projected database, denoted by SDB$|_\alpha$, is the collection of suffixes of sequences in SDB with regard to prefix α. Table 3 shows an example of the projected database considering "*a*" as prefix.

3 Sequential Mining Algorithms

There exist many data mining techniques such as classification, clustering, association rule mining and others. This chapter focuses on sequential mining algorithms. We present the state of the art of recent algorithms, elaborating a classification based on their main objectives and principles. Thus, this classification divides the algorithms into three primary types: SP Mining, Sequential Rule Mining and Sequence Prediction. Each of these can be split into different criteria and strategies.

3.1 Sequential Pattern Mining

3.1.1 Frequent Sequential Pattern Mining

It consists of finding subsequences appearing frequently in a set of sequences called sequential pattern or frequent subsequence. The frequency of these patterns is no less than a minimum support threshold *minsup* specified by the user. The common frequent Sequential Pattern Mining algorithms are:

The Generalized Sequential Patterns (GSP) algorithm [11] is an Apriori-like method and was one of the first algorithms that studied SPs after Apriori-All. The database is scanned multiple times. The first pass determines the support of each item, which is the number of data sequences that include the item. It simply means counting the occurrences of singleton transactions (containing one element) in the given database (one scan of the whole database). After this process, nonfrequent items are removed, and each transaction consists now of its original frequent items. This result will be the input of the GSP algorithm. Like Apriori, GSP algorithm makes multiple database scans. At the first pass, all single items of length 1 sequences (1-sequences) are counted. At the second pass, frequent 1-sequences are used to define the sets of candidate 2-sequences, and another scan is made to calculate their support. Same process is used to discover the candidate 3-sequences but using frequent 2-sequences, and so on until no more frequent sequences are found. GSP algorithm is composed of two techniques:

1. Candidate Generation: Only candidates with minimum support or above are conserved until no new candidates are found. This technique generates an enormous number of candidate sequences and then tests each one with respect of the user-defined *minsup*.

   After the first scan of the database and obtaining frequent $(k-1)$-frequent sequences $F(k-1)$, a joining procedure of $F(k-1)$ with itself is made and any infrequent sequence is pruned if at least one of its subsequences is not frequent.
2. Support Counting: a hash tree-based search is used. Finally nonmaximal frequent sequences are removed.

The GSP algorithm also allows frequent sequences discovery with time constraints. It can calculate the difference between the end-time of the element just found and the start-time of the previous element. This time is user defined and called maximum and minimum gap. Furthermore, it supports the concept of a sliding window (defines the interval of time between items in the same transaction).

The Sequential PAttern Discovery using Equivalence classes, SPADE [10] is based on a vertical id-list database format in which each sequence is associated to a list of items in which it appears: each subsequence is originally associated to its occurrence list. The frequent sequences can be found by using the intersection on id-lists. The size of the id-lists is the number of sequences in which an item appears. SPADE reduces the search space by aggregating SPs into equivalent classes and thus reduces the execution time. Thereby, two k-length sequences are in the same equivalence class if they share the same k-1 length prefix.

In his first step, SPADE computes the support of length 1 sequences, and this is done in a single database scan. In its second step, SPADE computes the support of 2-sequences and this is done by transforming the vertical representation into a horizontal representation in memory. This counting process is done with one scan of data and uses a bi-dimensional matrix. The idea consists of joining $(n-1)$ sequences using their id-lists to obtain n-subsequences. If the size of id-list is greater than *minsup*, then the sequence is frequent. The algorithm can use a breadth-first or a depth-first search method for finding new sequences. The algorithm stops when no more frequent sequences are found.

Sequential PAttern Mining, SPAM [12], is a memory-based algorithm and uses vector of bytes (bitmap representation) to study the existence (1) or absence (0) of an item in a sequence after loading the database into the memory. Candidates are generated in a tree by an S-extension that adds an item in another transaction, and by an I-extension that appends the item in the same transaction. The candidates are verified by counting the bytes with a value of one with the defined *minsup*.

The algorithm is efficient for mining long sequential patterns. Depth-first search is used to generate candidate sequences, and various I-step pruning and s-Step pruning are used to reduce the search space.

The transactional data are stored using a vertical bitmap representation, which allows for efficient support counting as well as significant bitmap compression. One new feature introduced with SPAM is that it incrementally outputs new frequent itemsets in an online fashion.

The Prefix-projected Sequential Pattern Mining, known as PrefixSpan [13], is a pattern-growth-based algorithm that discovers SPs using the idea of projected database. The algorithm studies the prefix subsequences instead of exploring all the possible occurrences of frequent subsequences (refer to the definitions 16 and 17). Then, it performs a projection on their corresponding post-fix subsequences. Frequent sequences will grow by mining only local frequent patterns, showing the efficiency of this algorithm.

The Last Position Induction algorithm (LAPIN) [14] is used for the extraction of long sequences and the reduction of the search space. It uses a lexicographical tree as the search path with DFS strategy. LAPIN-LCI procedure tests each item in the local candidate list and directly decides whether the item can be added to the prefix sequence or not. It compares the item's last position with the prefix border position. The algorithm assumes that the last position of an item i is helpful to decide whether this item could be appended to a frequent sequence of length k in order to get a frequent sequence of $k + 1$ length.

The CM-SPAM and CM-SPADE [15] are extensions of the two well-known algorithms SPADE and SPAM to which is added a new structure called Co-Occurrence MAP (C-MAP). The latter is used to store co-occurrence information by dividing them into CMAPi and CMAPs substructures. The first stores the items that succeed each item by i-extension and the second stores the items that succeed each item by s-extension at least *minsup* times. Let S be the sequence $\{I_1, I_2, \ldots,$

$I_n\}$. An item k is said to succeed by i-extension to an item j in S, iff j and $k \in I_x$ for an integer x such that $1 \le x \le n$ and $k >_{\text{lex}} j$. An item k is said to succeed by s-extension to an item j in S, iff $j \in I_v$ and $k \in I_w$ for some integers v and w such that $1 \le v < w \le n$.

The i-extension of pattern P with an item x is considered nonfrequent if there exists an item i in the last itemset of P such that (i,x) is not in CMAPi. Same for the pruning of s-extension: The s-extension of a pattern P with an item x is infrequent if there exists an item i in P such that (i, x) is not in CMAPs.

3.1.2 Closed Sequential Pattern Mining

A Closed Sequential Pattern (CSP) is not necessarily included in another pattern having the same support. The set of CSPs is much smaller than the set of SPs making mining more efficient. There exists no super-pattern S′ of pattern S having the same support of S. Then S is a closed sequential pattern; in other words, Closed Pattern Mining means that for the same support the mining process will mine the longest pattern. Common Sequential Patterns algorithms are given in the following.

The CloSpan algorithm [16] is based on mining frequent closed sequences in large data sets instead of exploring all frequent sequences and is used to mine long sequences. Its main advantage is in time and space reduction. The algorithm is divided into two stages. In the first, it generates a set of all frequent sequences and eliminates the nonclosed sequences in the second. It represents data with lexicographical tree or order.

A Lexicographic Sequence Tree (LST) can be constructed as follows:

1. Each node in the tree corresponds to a sequence, and the root is a null sequence.
2. If a parent node corresponds to a sequence S_1, its child is either an itemset-extension of S_1, or a sequence-extension of S_1.
3. The left sibling is less than the right sibling in sequence lexicographic order.

The BI-Directional Extension (BIDE+) [17] is an extension of the BIDE algorithm that mines closed SPs and avoids problem of the candidate maintenance-and-test paradigm used by CloSpan. It works in a DFS manner in order to generate the frequent closed patterns and consumes less memory compared to the previous version.

The ClaSP [18] is based on the SPADE algorithm and was the first to mine closed frequent SPs in vertical databases. ClaSP has two phases: The first one generates a subset of frequent sequences called Frequent Closed Candidates (FCC), which is kept in main memory; and the second step executes a post-pruning phase to eliminate all nonclosed sequences from FCC to finally obtain exactly FCS.

CM-ClaSP [9] is an extension of ClaSP based on the new representation of data called C-MAP as discussed in CM-SPADE and CM-SPAM.

3.1.3 Maximal Sequential Pattern Mining

Sequential Pattern Mining may return too many results, making it difficult for the user to understand and analyze. Mining maximal SPs may be a solution. A Maximal SP is a pattern that is not included in another pattern. Maximal Pattern Mining algorithms are presented in the following.

The MaxSP [19] is inspired by the PrefixSpan algorithm. It is based on a pattern-growth algorithm that aims to extract maximal SPs without maintaining candidates. It has an integrated BIDE-like mechanism that checks if a pattern is maximal. MaxSp reduces the redundancy in SPs that could be time consuming and requires a lot of storage space.

The Vertical Maximal Sequence Patterns (VMSP) [20] is based on the SPAM search procedure that generates the pattern and explores candidate patterns having same prefix in a recursive manner. VMSP integrates three strategies: Efficient Filtering of Nonmaximal Patterns (EFN), Forward Maximal Extension Checking (FME) and Candidate Pruning by Co-Occurrence Map (CPC).

3.1.4 Compressing Sequential Pattern Mining

This kind of algorithm is used to reduce redundancy and thus to minimize the size of mining results.

GoKrimp and SeqKrimp [21] are two compressing SPs mining algorithms, based on the Krimp algorithm. They explore directly compressing patterns and avoid the resource-consuming candidate generation. SeqKrimp uses a frequent closed SPs mining algorithm to generate a set of candidate patterns. It gets the candidate pattern set and returns a good subset of compressing patterns, then greedily calculates the benefits of adding/extending a given pattern from the candidates. This procedure is repeated until no more useful patterns can be added. GoKrimp uses the same procedures but is an ameliorated version of SeqKrimp. It searches for a set of sequential patterns that compresses the data most based on the minimum description length principle; informally, the best model is the one that compresses the data the most. What differentiates GoKrimp is that it is parameter free. Users are not supposed to set a minimum support, which is a difficult decision in some cases. A dependency test is provided to consider only related patterns to extend a given pattern. This technique aims to avoid the excessive tests of all possible extensions and makes the GoKrimp faster than SeqKrimp.

3.1.5 Top-K Sequential Pattern Mining

In SP mining algorithm, tuning the *minsup* parameter to get enough patterns is a difficult and time-consuming task. To remedy this issue, Top-K Sequential Pattern mining algorithms were implemented to return k SPs.

TSP (Top-K Closed Sequential Patterns) [22] uses the concept of pattern-growth and projection-based SP mining of PrefixSpan algorithm, and then performs a multi-pass mining to find and grow patterns. After closed pattern verification phase, the algorithm applies the minimum length constraint verification, which reduces the search space.

TKS (Top-K Sequential Patterns) [23] uses a vertical bitmap database representation. It adapts the SPAM search procedure to explore the search space of patterns to transform it to a Top-K algorithm. Then, TSK extends the most promising patterns, meaning that it finds patterns with high support in an early stage and discards infrequent items. Finally, the algorithm uses a PMAP (Precedence MAP) data structure to prune the search space.

3.2 Sequential Rules Mining

3.2.1 Sequential Rules

Mining frequent patterns is not sufficient in decision-making. Sequential rules mining and sequence prediction (next section) are necessary. A sequential rule indicates that if some item(s) occur in a sequence, some other item(s) are likely to occur afterward with a given confidence or probability. Common Sequential Rules mining algorithms are presented in the following.

CMDeo [24] was first designed to explore rules in a single sequence. It explores the search space in BFS and extracts all valid rules of size 1*1 respecting minimum support and confidence. Similarly to Apriori, CMDeo generates a huge amount of valid rules by applying a left and a right expansion.

RuleGrowth [25] explores sequential rules for several sequences and not only for one. It is based on the pattern-growth approach in finding the sequential relations that explores rules between two items and expands them left and right.

CMRules [8] is an alternative of CMDeo. It searches for the association rules to reduce the search space, then it removes the rules that do not respect minimum support and confidence. Therefore, it could be used to discover both sequential rules and association rules at the same time.

The Equivalence class-based sequential Rule Miner (ERMiner) [26] algorithm uses a vertical representation to avoid database projection. It mines the search space through equivalence classes to generate rules with the same antecedent or consequent.

3.2.2 Top-K Sequential Rules Mining

Specifying the number of sequential rules to be found may overcome the difficulty in fine-tuning sequential rules parameters like *minsup* and *minconf*. The idea of discovering Top-K nonredundant rules comes after the difficulty and the time-consuming task to tune the minimum support value by the user. Moreover, the sequential rule mining algorithms usually return a high level of redundancy. To solve both problems, the Top-K Sequential Rules Mining algorithms let the user indicate k, which is the number of rules to be discovered.

The TopSeqRule [27] was the first to address the Top-K sequential rules mining. It generates rules for several sequences based on the RuleGrowth search strategy integrated with the general process for mining Top-K patterns. To optimize results, it first generates the most promising rules and reduces the search space by increasing *minsup*.

Top-K Nonredundant Sequential Rules TNS [28] is used to discover the Top-K non-redundant sequential rules. It adopts the TopSeqRule to mine the Top-K rules and adapts it to eliminate redundancy. The algorithm gives an approximation and thus does not guarantee to retrieve the Top-K nonredundant rules. TNS has a positive integer parameter called delta to increase the result's exactitude. Results are more exact with a higher value of delta.

3.2.3 Sequential Rules with Window Size Constraints

This kind of algorithm returns all sequential rules with regard to the specified *minsup* and *minconf* appearing within a window size.

TRuleGrowth [29] is an extension of the RuleGrowth with a sliding window constraint. It is very useful in the discovery of temporal patterns (patterns that happen within a maximum time interval). TRuleGrowth allows the user to specify other optional parameters like the minimum antecedent length and the maximum consequent length. These parameters define respectively the minimum number of items appearing in the left side and the maximum number of items in the right side of a rule, knowing that the left side is the antecedent and the right side is the consequent.

3.2.4 Sequence Prediction

In many applications, it is very important to predict the next element in a sequence. Given a set of sequences, the idea is to predict the next element in a sequence S, based on a set of training sequences. Various applications use the sequence prediction algorithms. For example, one may need to know the next web page to be visited by the user, based on his/her, and/or other users' histories.

The Compact Prediction Tree (CPT) [30] is a lossless sequence prediction model that uses all information in the sequence for prediction. It consists of two phases: the training phase and the prediction phase. The first compresses the sequences in a prediction tree. A given sequence S is predicted by finding all sequences that contain the last x items from S in any order and in any position. CPT is more efficient than other existent algorithms such as Prediction by Partial Matching (PPM) [31], Dependency Graph (DG) [32] and All-K-th-Order Markov [33].

CPT+ [34] is an enhanced version of CPT where Frequent Subsequence Compression (FSC), Simple Branch Compression (SBC) and Prediction with improved Noise Reduction (PNR) strategies were added to improve prediction time and precision.

4 Conclusion

This chapter summarizes the most recent and common algorithms on the sequential mining paradigm. It does not aim to give a deep explanation about each algorithm, but it mentions its purpose and gives an idea about how it works. One should refer to the related article of each algorithm for additional details. For further explanation and ease of understanding, this chapter also presents a classification for the sequential mining algorithms. They are arranged by their usage. This classification was based on three main axes: frequent sequential pattern mining, sequential rules mining and sequence prediction. Important terms and notations in the data mining domain were first introduced. Then, a short definition introduced each class to let the reader have a quick idea about it. Later, the most important and recent algorithms in each axis were investigated with a brief description about their methods and implementations.

5 Discussion

The diagram in Fig. 1 consists of a classification tree containing the most recent algorithms and their extensions. This tree can help researchers in choosing the appropriate algorithm according to their needs especially when it comes to sequential pattern mining. Sequential mining is efficient for applications that are time-based or take into consideration the order of the event. Sequential mining has proven its efficiency through time in the economic field starting from GSP that analyzes the transactions of customers in order to improve the income and marketing strategies. Sequential Mining started showing its importance in medical field, making it a very promising field for researchers and programmers.

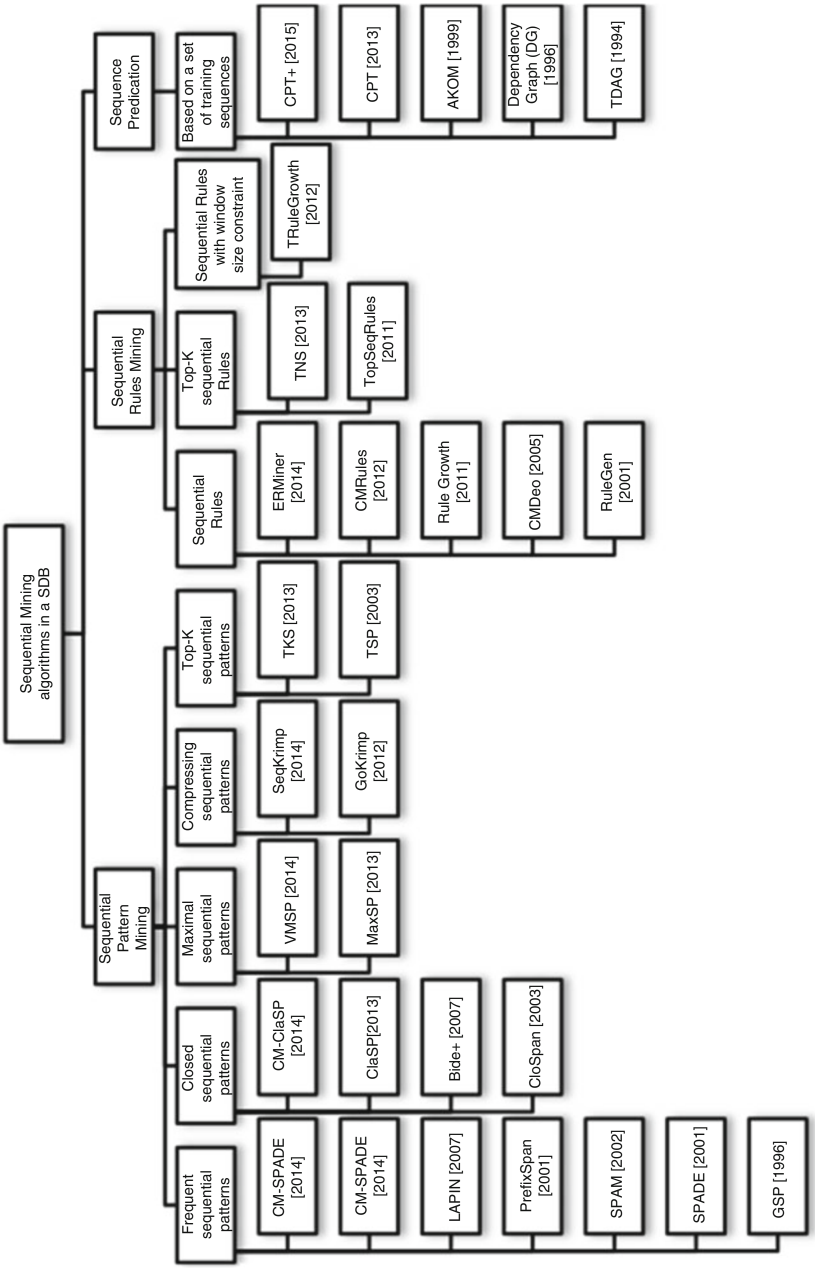


Fig. 1 Classification of sequential mining algorithms

References

1. A.P. Wright, A.T. Wright, A.B. McCoy and D.F. Sittig, “The use of sequential pattern mining to predict next prescribed medications”. Journal of biomedical informatics, 53, pp.73–80, 2015.
2. G. Bruno and P. Garza, “Temporal pattern mining for medical applications”. Data Mining: Foundations and Intelligent Paradigms, pp.9–18. 2012
3. K. Uragaki, T. Hosaka, Y. Arahori, M. Kushima, T. Yamazaki, K. Araki and H. Yokota, “Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines”. In IEEE Symposium on Computers and Communication (ISCC), (pp. 20–25). IEEE. 2016.
4. K. Choi, S. Chung, H. Rhee and Y. Suh, Classification and sequential pattern analysis for improving managerial efficiency and providing better medical service in public healthcare centers. Healthcare informatics research, 16(2), pp.67–76, 2010.
5. C. Bou Rjeily, G. Badr, A. Hajjam El Hassani and E. Andres, “Sequence Prediction Algorithm for Heart Failure Prediction”, International Conference e-Health, ISBN: 978-989-8533-65-4, pp.109–116, 2017.
6. C. Bou Rjeily, G. Badr, A. Hajjam El Hassani and E. Andres, “Predicting Heart Failure Class using a Sequence Prediction Algorithm”, Fourth International Conference on Advances in Biomedical Engineering (ICABME), 2017
7. R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases”, In ACM sigmod record(Vol. 22, No. 2, pp. 207–216). ACM, 1993.
8. P. Fournier-Viger, U. Faghihi, R. Nkambou, E. Mephu Nguifo, “CMRules: Mining Sequential Rules Common to Several Sequences. Knowledge-based Systems”, Elsevier, 25(1): 63–76, 2012.
9. J. Han, J. Pei, Y. Yin, and R. Mao, “Mining frequent patterns without candidate generation: A frequent-pattern tree approach”, Data mining and knowledge discovery, 8(1), pp.53–87, 2000.
10. M.J. Zaki, “SPADE: An efficient algorithm for mining frequent sequences”, Machine learning, 42(1–2), pp.31–60, 2001.
11. R. Srikant, and R. Agrawal, “Mining sequential patterns: Generalizations and performance improvements”, In International Conference on Extending Database Technology (pp.1–17). Springer Berlin Heidelberg, 1996.
12. J. Ayres, J. Flannick, J. Gehrke, J. and T. Yiu, “Sequential pattern mining using a bitmap representation”, In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining(pp. 429–435). ACM, 2002.
13. J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu, “Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth”, In proceedings of the 17th international conference on data engineering, pp. 215–224, 2001.
14. Z. Yang, Y. Wang, and M. Kitsuregawa, M., “LAPIN: effective sequential pattern mining algorithms by last position induction for dense databases”, In International Conference on Database systems for advanced applications (pp. 1020–1023). Springer Berlin Heidelberg, 2007.
15. P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas, “Fast vertical mining of sequential patterns using co-occurrence information”, In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 40–52). Springer International Publishing, 2014.
16. X. Yan, J. Han, R. Afshar R., “CloSpan: Mining Closed Sequential Patterns in Large Datasets”, Proceedings of the 2003 SIAM International Conference on Data Mining, 2003.
17. J. Wang, and J. Han, “BIDE: Efficient mining of frequent closed sequences”, In Data Engineering, 2004. Proceedings. 20th International Conference on (pp. 79-90). IEEE, 2004.
18. A. Gomariz, M. Campos, R. Marin, and B. Goethals, “Clasp: An efficient algorithm for mining frequent closed sequences” In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 50-61). Springer Berlin Heidelberg, 2013.

19. P. Fournier-Viger, C.W. Wu, and V.S. Tseng, “Mining maximal sequential patterns without candidate maintenance”, In International Conference on Advanced Data Mining and Applications (pp. 169-180). Springer Berlin Heidelberg, 2013.
20. P. Fournier-Viger, C.W. Wu, A. Gomariz, and V.S. Tseng, “VMSP: Efficient vertical mining of maximal sequential patterns”, In Canadian Conference on Artificial Intelligence (pp. 83-94). Springer International Publishing, 2014.
21. H.T. Lam, F. Mörchen, D. Fradkin, and T. Calders, “Mining compressing sequential patterns”, Statistical Analysis and Data Mining, 7(1), pp.34-52, 2014.
22. P. Tzvetkov, X. Yan, and J. Han, “TSP: Mining Top-k Closed Sequential Patterns”, Knowledge and Information Systems, vol. 7, no. 4, pp. 438-457, 2005.
23. P. Fournier-Viger, A. Gomariz, T. Gueniche, E. Mwamikazi, and R. Thomas, “TKS: efficient mining of top-k sequential patterns”, In International Conference on Advanced Data Mining and Applications (pp. 109-120). Springer Berlin Heidelberg, 2013.
24. J. Deogun, and L. Jiang, “Prediction mining–an approach to mining association rules for prediction”, In International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing (pp. 98-108). Springer Berlin Heidelberg, 2005.
25. P. Fournier-Viger, R. Nkambou, and V.S.M. Tseng, “RuleGrowth: mining sequential rules common to several sequences by pattern-growth”, In Proceedings of the 2011 ACM symposium on applied computing (pp. 956-961), 2011.
26. P. Fournier-Viger, T. Gueniche, S. Zida, and V.S. Tseng, “ERMiner: sequential rule mining using equivalence classes”, In International Symposium on Intelligent Data Analysis (pp. 108-119). Springer International Publishing, 2014.
27. P. Fournier-Viger, and V.S. Tseng, “Mining top-k sequential rules”, In International Conference on Advanced Data Mining and Applications (pp. 180-194). Springer Berlin Heidelberg, 2011.
28. P. Fournier-Viger, and V. S Tseng, “TNS: mining top-k non-redundant sequential rules”, In Proceedings of the 28th Annual ACM Symposium on Applied Computing, 2013.
29. P. Fournier-Viger, C.W. Wu, V.S. Tseng, and R. Nkambou, “Mining sequential rules common to several sequences with the window size constraint”, In Canadian Conference on Artificial Intelligence (pp. 299-304). Springer Berlin Heidelberg, 2012.
30. T. Gueniche, P. Fournier-Viger, and V.S. Tseng, “Compact prediction tree: A lossless model for accurate sequence prediction”, In International Conference on Advanced Data Mining and Applications (pp. 177-188). Springer Berlin Heidelberg, 2013.
31. J. Cleary, I. Witten, “Data compression using adaptive coding and partial string matching”, IEEE Trans. on Inform. Theory, vol. 24, no. 4, pp. 413-421, 1984.
32. V. N, Padmanabhan, J.C. Mogul, “Using Prefetching to Improve World Wide Web Latency”, Computer Communications, vol. 16, pp. 358-368, 1998.
33. J. Pitkow, P. Pirolli, “Mining longest repeating subsequence to predict world wide web surfing”, In: USENIX Symposium on Internet Technologies and Systems, Boulder, CO, pp. 13-25, 1999.
34. T. Gueniche, P. Fournier-Viger, R. Raman, and V.S. Tseng, “CPT+: Decreasing the time/space complexity of the Compact Prediction Tree”, In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 625-636). Springer International Publishing, 2015.

Object Detection Based on CNNs: Current and Future Directions

Long Chen, Abdul Hamid Sadka, Junyu Dong, and Huiyu Zhou

1 Introduction

The goal of object detection is to learn a visual model for concepts such as cars and use this model to localize these concepts in an image. As shown in Fig. 1, given an image, object detection aims at predicting the bounding box and the label of each object from the defined classes in the image. This requires the ability to robustly model invariants against illumination changes, deformations, occlusions and other intra-class variations. Among a number of vision tasks, object detection is one of the fastest moving areas due to its wide applications in surveillance [1, 2] and autonomous driving [3, 4].

L. Chen · H. Zhou
School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK
e-mail: lchen15@qub.ac.uk; h.zhou@ecit.qub.ac.uk

A. H. Sadka (✉)
Department of Electronic and Computer Engineering, Brunel University, London, UK
e-mail: abdul.sadka@brunel.ac.uk

J. Dong
Department of Computer Science and Technology, Ocean University of China, Qingdao, China
e-mail: dongjunyu@ouc.edu.cn

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_2

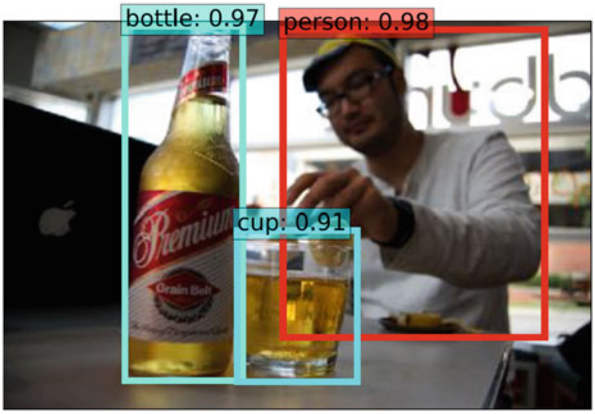


Fig. 1 Bounding boxes and labels with corresponding class probabilities predicted by detectors

2 From Handcrafted Features to Deep CNNs Methods

2.1 Handcrafted Features

Before deep CNNs, convolutional neural networks [5], were introduced, the progress on various visual recognition tasks had been considerably based on the use of handcrafted features, such as SIFT [6] and HOG [7]. Handcrafted features can be broadly divided into three categories:

1. Interest Point Detection. These methods use certain criteria to select pixels, edges and corners as well-defined local texture features. Among them, Sobel, Prewitt, Roberts, Canny and LoG (Laplacian of Gaussian) are typical edge detection operators [8–11], while Harris, FAST (Features from Accelerated Segment Test), CSS (Curvature Scale Space) and DOG (Difference of Gaussian) are typical corner detection operators [6, 12, 13]. Interest point detection methods usually have a certain geometric invariance which can be found at a small computational cost.
2. Methods based on local features. These methods mainly extract local features, which are different from global features such as colour histograms, which are ideal for dealing with partial occlusion of target objects. Commonly used local features include Scale-Invariant Feature Transform (SIFT) [6], HOG (Histogram of oriented gradient) [7], Haar-like [14] and Local Binary Pattern [15, 16]. Local features are informative, unique, with strong invariance and distinguishability. But the calculation is generally complicated, and local features are further developed to have better representations in recent years.
3. Methods based on multi-feature combination. A combination of interest point and local feature extraction methods can be used to handle the deficiency of using a single feature to represent target objects. DPM (Deformable Part-based

Model) [17] is an effective multi-feature combination model which has been widely applied to the object detection task and has achieved good performance, such as pedestrian detection [14, 16], face detection [15, 18] and human pose estimation [19]. In [20], three prohibitive steps in the cascade version of DPM were accelerated, which greatly improved the detection speed.

The characteristics of handcrafted features are largely dependent on experience and environments, where most of the test and adjustment workloads are undertaken by the user, which is time-consuming. In contrast, an important viewpoint in the deep learning theory, which has drawn much attention in recent years, is that handcrafted descriptors, as the first step in a visual system, tend to lose useful information. Directly learning task-related feature representation from raw images is more effective than handcrafted features [21].

For object detection tasks, handcrafted features based systems have become a dominant paradigm in the literature before deep CNNs were introduced. If we look at system performance on the canonical visual recognition task, PASCAL VOC object detection [22], it is acknowledged that certain progress has been made during 2010–2012, by building ensemble systems and employing variants of successful methods. Recently, Convolutional Neural Networks (CNNs) [5] have produced impressive performance improvements in many computer vision tasks since 2012, such as image classification, object detection and image segmentation. CNNs witnessed its frequent use in the 1990s (e.g., [5]), but then became less used, particularly in computer vision, with the powerful impact of support vector machines (SVMs) [23]. In 2012, Krizhevsky et al. [24] rekindled interests in CNNs by showing substantially high image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [25]. Their success resulted from training a large CNN on 1.2 million labelled images, together with a few twists on [5] (e.g., ‘dropout’ regularization). The significance of deep CNNs methods will be introduced in the following section.

2.2 Deep Learning Approaches

Convolutional Neural Networks [5] is the first successful method in deep learning approaches. The key difference between CNNs-based and conventional approaches is that in the former, the feature representation is learned instead of being designed by the user. These recent successes were built upon the powerful deep features that are learned from large-scale datasets, which accompany accurate annotations with the drawback that a large number of training samples are required for training the classifier. Among many variants of the CNNs-based approaches, they can be roughly divided into two streams: region proposal-based methods and proposal-free methods.

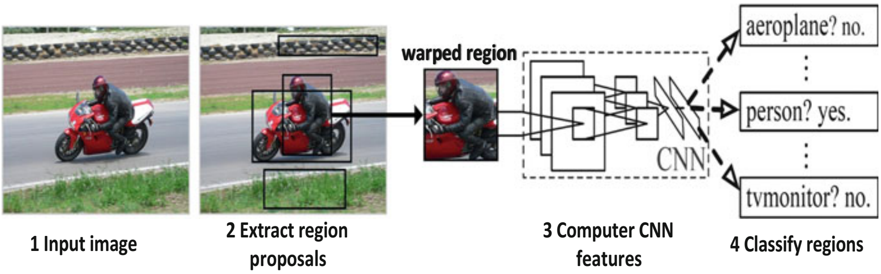


Fig. 2 The overview of the R-CNN detection system. (1) Input an image, (2) extracts region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs

2.2.1 Region Proposal-Based Methods

The dominant paradigm in modern object detection is the region proposal-based method. The pioneering work Selective Search [26] consists of two stages: The first stage generates a sparse set of candidate proposals that should contain all the objects while filtering out the majority of negative locations and the second stage classifies the proposals into foreground or background. R-CNN [27] upgrades the second-stage classifier to a convolutional network yielding large gains in accuracy and ushering in the modern era of object detection (shown in Fig. 2). R-CNN requires high computational costs while each proposal is processed by the CNNs separately. Fast R-CNN [28] improved efficiency by sharing computation and using RoI (Region of Interest) pooling [29] to efficiently generate features for object proposals. Region Proposal Networks (RPNs) integrate proposal generation with the second-stage classifier in a single convolution network, forming the Faster RCNN framework [30]. R-FCN [31] further improved efficiency and accuracy by removing fully connected layers while adopting position-sensitive score maps for the final detection. However, one problem with the region-based methods is that in order to process a large number of proposals, the computation in the second stage is usually costly. To accelerate the detection process, proposal-free methods have been proposed for real-time detection.

2.2.2 Proposal-Free Methods

Proposal-free methods aim to eliminate the region proposal stage and directly train a single-stage end-to-end detector. Without the region proposal stage, they have the potential to be faster and simpler, but have trailed the accuracy of two-stage detectors thus far. YOLO [32] used a single feed-forward convolutional network to directly predict object classes and locations. Compared with region-based methods, YOLO no longer requires a second per-region classification operation so that it is extremely fast. SSD [33] improved YOLO in several aspects, including (1) using small convolutional filters to predict categories and anchor offsets for bounding box locations; (2) using pyramid features for prediction at different scales; (3) using default boxes and aspect ratios for adjusting varying object shapes. Those clever designs save considerable amounts of computation and perform much faster than Faster RCNN. The proposal-free detectors are usually easier to train with less computational efforts. However, such advantage is largely overwritten when the models are evaluated in benchmarks considering mean average precision (mAP) for high intersection-over-union (IoU) thresholds (e.g., KITTI car) since the two-stage methods are usually advantageous in performance. It achieved good results in datasets for the IoU threshold of 0.5. However, the performance drops significantly when we increase the bar for detection quality.

2.2.3 Fine-Tuning Strategy

When training supervised classifiers, we expect that there are sufficient labelled samples available for the target classes [1]. However, this requirement seems too demanding in some real-world applications. For example, many objects 'in the wild' follow a long-tailed distribution such that they do not occur frequently enough to collect and label a large set of representative exemplars to build the corresponding recognizers [34]. In addition, the labelling effort for many objects can be very expensive because the expert knowledge is required, for example, fine-grained bird recognition [35]. Under these circumstances, it is always expected to train effective classifiers with as few labelled samples as possible. Fine-tuning is one of the widely adopted paradigms to save efforts for labelling data in a supervised learning. It involves learning a generic feature representation on a large dataset of labelled images, and then specializing or fine-tuning the learned generic feature representation for a specific task at hand. Especially, in order to achieve good performance, most of the advanced object detection systems fine-tune classification networks that start from generic features learned on the ImageNet dataset using over a million labelled images and then specialize them for object detection tasks. Other approaches [36, 37] design specific backbone network structures for object detection, but still require pre-training the networks on the ImageNet classification dataset.

Fine-tuning object detectors from the pre-trained classification models has at least two advantages. First, there are many state-of-the-art deep models publicly available. It is convenient to reuse them for object detection. Second, fine-tuning can quickly generate the final model and requires much less instance-level annotated training data than the classification task. However, there are also critical limitations when adopting the pre-trained networks in object detection: (1) A limited structure design space. The pre-trained network models are mostly from the ImageNet-based classification task, which are usually very heavy—containing a huge number of parameters. (2) Learning bias. As both the loss functions and the category distributions between classification and detection tasks are different, this will lead to different searching/optimization spaces. Therefore, learning may be biased towards a local minimum, which is not the best for the detection task. Model fine-tuning for the detection task can alleviate this bias to some extent but not fundamentally. (3) Domain mismatch. As known, fine-tuning can mitigate the gap due to different target category distributions. However, it is still a severe problem when the source domain (ImageNet) has a huge mismatch to the target domain such as depth images.

Finally, it is worth noting that some recent work attempts to train CNNs from scratch. The proposed approach has very appealing advantages over the existing pre-training solutions [38, 39]. In semantic segmentation, Jégou et al. [40] demonstrated that a well-designed network structure can outperform state-of-the-art solutions without using the pre-trained models. It extends DenseNets [39] to fully convolutional networks by adding an up-sampling path to recover the original resolution. Shen et al. [41] presented the Deeply Supervised Object Detector (DSOD), a framework that can learn object detectors from scratch and contribute to a set of design principles for training object detectors from scratch. Training CNNs from scratch is a promising future direction due to its wide applications, though not much work has been done in this area yet.

3 Current Research Directions

In the present section, we discuss current research directions. Current research has been focused on three principal directions for developing better object detection systems. The first direction relies on innovating the base architecture of the existing networks. It has been shown that using all the examples does not always lead to an optimal solution [42] and data selection is the key. So, another research direction focuses on how to better exploit the data itself. The third area of research is to use contextual reasoning, as it can be a rich source of information about an object identity, location and scale [31].

3.1 Excellent Base Architectures

Many innovative CNN structures have been proposed [24, 38, 43, 44]. Meanwhile, several regularization techniques have also been proposed to further enhance the model capabilities. Krizhevsky et al. proposed a new convolution neural network AlexNet [24], followed by a series of improved models, such as ZFNet [43], VGG [37], GoogLeNet [44] and ResNet [38], proposed by other researchers. Table 1 shows the performance comparison of the classical CNN model in the image classification task of ILSVRC. The error rate in the image classification task of ILSVRC is reduced every year. The image classification top-5 error rate is getting lower as the base architecture becomes increasingly deeper. Although these network architectures are designed for image classification tasks, people aim to

Table 1 Performance comparison of the classical CNN model in image classification task of ILSVRC

CNN architecture	Top-5 error rate (%)
AlexNet [24]	16.4
ZFNet [43]	14.8
VGG [37]	7.3
GoogLeNet [44]	6.7
ResNet [38]	3.57
Inception-ResNet-v2 [45]	3.08

solve one of the most fundamental questions—how to create more powerful feature representation.

Given those CNN models which have strong feature representation, applying them to the target detection task results in good detection accuracy. He et al. [38] proposed residual learning blocks with skip connections, which enable training very deep detection networks with more than 100 layers. Huang et al. [39] proposed DenseNets with dense layer-wise connections. Kim et al. [46] proposed PVANet for object detection, which consists of the simplified 'Inception' block from GoogleNet. Huang et al. [47] investigated various combinations of network structures and detection frameworks, and found that Faster R-CNN with Inception-ResNet-v2 [45] achieved the best performance. Lin et al. [48] designed a simple one-stage object detector called RetinaNet, named for its dense sampling of object locations in an input image. Its design features include an efficient in-network feature pyramid and the use of anchor boxes. Thanks to these excellent network structures, the accuracy of the object detection task has been greatly improved. Performance comparison of some object detection methods on public datasets can be seen in Table 2.

3.2 Hard Example Mining

Training data plays a critical role in machine learning. The data selection strategy along the training process could significantly impact the performance of the learned model. For detection datasets which contain an overwhelming number of easy examples and a small number of hard examples, automatic selection of these hard examples can make training more effective and efficient. Hard example mining is one technique of allowing the learning system to select the most informative samples to train the model. The underlying assumption in hard example mining is that the samples have different information and only a small portion of the samples can provide sufficient information for supervised learning. In fact, the information of each sample is different; therefore, if the most representative/informative samples are selected and labelled, even a few labelled samples can provide sufficient knowledge to construct effective classifiers. Hard example mining has existed for at least 20 years, which was first introduced in [49] in the mid-1990s (if not earlier) for training face detection models. Their key idea is to perform training on a sparse set of hard examples and prevent the vast number of easy negatives from overwhelming

Table 2 Performance comparison of some object detection methods on public datasets

Datasets	Methods	mAP(%)
VOC2007	Fast R-CNN (VGG16)	70.00
	Faster R-CNN (VGG16)	73.20
	Faster R-CNN (VGG16)[a]	78.80
	Faster R-CNN (ResNet)	76.40
	Faster R-CNN (ResNet)[a]	85.60
	YOLO	63.40
	YOLOv2(544 × 544)	78.60
	SSD300 (VGG16)	72.10
	SSD500 (VGG16)	75.10
	ION	79.20
	HyperNet (VGG16)	76.30
	R-FCN (ResNet-101)	79.50
	R-FCN (ResNet-101)[a]	83.60
	PVANET	83.80
VOC2012	Fast R-CNN (VGG16)	68.40
	Faster R-CNN (VGG16)	70.40
	Faster R-CNN (VGG16)[a]	75.90
	Faster R-CNN (ResNet)[a]	83.80
	YOLO	57.90
	YOLOv2(544 × 544)	73.40
	Fast R-CNN, YOLO	70.70
	SSD300 (VGG16)	70.30
	SSD300 (VGG16)[a]	79.30
	SSD500 (VGG16)	73.10
	SSD512 (VGG16)	78.50
	SSD512 (VGG16)[a]	82.20
	ION	76.40
	OHEM, Fast R-CNN (VGG16)[a]	80.10
	HyperNet (VGG16)	71.40
	R-FCN (ResNet-101)	77.60
	R-FCN (ResNet-101)[a]	85.00
	R-FCN, ResNet Ensemble[a]	88.40
	PVANET	82.50
	Faster R-CNN, PVANET[a]	84.20
MSCOCO2015(@[0.5–0.95])	Fast R-CNN (VGG16)	19.70
	Faster R-CNN (VGG16)[a]	21.90
	Faster R-CNN (ResNet)[a]	37.40
	SSD300 (VGG16)	20.80

(continued)

Table 2 (continued)

Datasets	Methods	mAP(%)
	SSD500 (VGG16)	24.40
	ION	33.10
	R-FCN (ResNet-101)	29.20
	R-FCN (ResNet-101)[a]	31.50
	YOLOv2	21.60
MSCOCO2015(@0.5)	Fast R-CNN (VGG16)	35.90
	Faster R-CNN (VGG16)[a]	42.70
	Faster R-CNN (ResNet)[a]	59.00
	SSD300 (VGG16)	38.00
	SSD500 (VGG16)	43.70
	ION	55.70
	R-FCN (ResNet-101)	51.50
	R-FCN (ResNet-101)[a]	53.20
	YOLOv2	44.00

For VOC2007 dataset, the training set is the union of VOC2007 trainval and VOC2012 trainval, the testing set is VOC2007 test; for VOC2012, the training set is the union of VOC2007 trainval, VOC2007 test and VOC2012 trainval, the testing set is VOC2012 test

[a] Indicates using the union of MS COCO dataset and PASCAL dataset as training set; @ [0.5–0.95] means AP (averaged precision over IoU thresholds between 0.5 and 0.95) defined in COCO metric

the detector during the training. This strategy leads to an iterative training algorithm that alternates between updating the detection model given the current set of examples, and then using the updated model to find new false positives to add to the training set. The process typically commences with a training set consisting of all the object examples and a small, random set of background examples.

Hard example mining has seen widespread use in object detection research. Hard example mining algorithms are commonly used when optimizing SVMs [17, 26, 27]. In this case, the training algorithm maintains a working set of examples and alternates between training an SVM on the working set, and updating the working set by removing some examples and adding others according to a specific rule [17]. The rule removes easy examples since they provide little information to update the current model. Conversely, the rule adds hard examples which can provide sufficient information to accelerate the network training. Applying this rule leads to a global SVM solution. Hard example mining has also been applied to a number of models including shallow neural networks [50], boosted decision trees [51] and deep CNNs [52–55]. In this kind, an algorithm usually starts with a dataset of positive examples and a random set of negative examples. The machine learning model is then trained on that dataset and subsequently applied to a larger dataset to harvest false positives. The false positives are then added to the training set and then the model is trained again.

3.3 Contextual Reasoning

Context is known to play an important role in visual recognition [56]. Using contextual reasoning, proxy tasks for reasoning and other top-down mechanisms can improve image representation for object detection. Sermanet et al. [57] used two contextual regions centred on each object for pedestrian detection. In [58], in addition to specific features, features from the entire image are used to improve region classification. He et al. [29] implemented context in a more implicit way by aggregating CNN features prior to classification using different sizes' pooling regions. More recently, [59] proposed to use ten contextual regions around each object with different crops. Shrivastava and Gupta [60] used segmentation as a way to contextually prime object detectors and provide feedback to initial layers. Bell et al. [61] used a skip network architecture and the features extracted from multiple layers of representation in conjunction with contextual reasoning. Other approaches include using top-down features for incorporating context and finer details [62, 63], which leads to improved detection results.

4 Open Problems and Future Directions

In the following, we outline the problems that we believe have not been addressed, or addressed only partially in the literature, and may become interesting and relevant research directions.

4.1 Scale Invariance

To handle different variations, such as occlusion and deformations, current CNNs-based classifiers and detectors usually use a data-driven strategy—collect large-scale datasets which have object instances under different conditions. For example, the COCO dataset [64] has more than 10K examples of cars under different occlusions and deformations. We hope that these examples capture all possible variations of a visual concept and the classifier can then effectively model invariances. For CNN-based object detectors, the variance in pose and appearance can be handled by the capacity of convolutional neural networks. However, the CNN does not inherently hold scale invariance.

In academic research, two techniques are introduced to address this problem: (1) Simple multi-scale testing on image pyramids can be used to avoid the problem and achieve good accuracy [27, 28, 30, 31]. However, multi-scale testing leads to heavy computational costs. (2) The second way is to fit a CNN model to multiple scales [33, 61, 65]. They either construct a stronger network structure by combining features from different depths of a network or directly predict objects at different depths of a network. These attempts have been, to some extent, successful under this kind of problem, but they may also lead to an increase in model size and computation. So, further improvement is still required.

4.2 High Localization Accuracy

In many real-world applications, robustly detecting objects with high localization accuracy, namely to predict the bounding box location with high Intersection over Union (IoU) is crucial to the quality of service. For instance, in vision-based robotic arm applications, the process of generating robust and accurate operations in picking up an object is highly dependent on the object localization accuracy. In advanced driver assistance systems (ADAS), accurately localizing cars and pedestrians is also closely related to the safety of the autonomous actions.

R-CNN and its variants challenge the problem using a classification approach, and they employ regression as a post-processing stage to refine the localization of the proposed bounding boxes. Najibi et al. [66] modelled object detection as finding a path from a fixed grid to boxes tightly surrounding the objects, and slacked the regression process to several iterations for the reason that one step regression cannot handle the nonlinearity of the coordinates of bounding boxes. Gidaris et al. [67] proposed a novel object localization methodology that is based on assigning probabilities related to the localization task. Those probabilities provide useful information regarding the location of the object inside the search region and they can be exploited in order to infer its boundaries with high accuracy. Further improvements are required considering its importance in many practical applications.

4.3 Long-Tail Distribution

The ImageNet image classification dataset is a well-compiled dataset, in which objects of different classes have similar numbers of samples. In real applications, however, we will experience the long-tail distributions, where a small number of object classes appear very often but the others appear rarely. For object detection, some object classes such as persons have much more samples than the other object classes like sheep for both PASCAL VOC [22] and ImageNet [68] object detection datasets, as shown in Fig. 3. For deeply learned features, however, the feature learning will be dominated by the object classes with a large number of samples and the features are not good for object classes with fewer samples in the long tail. Therefore, the extreme class imbalance encountered during the training of detectors cannot learn discriminative features well for each category. Besides, the existence of many background samples makes the feature representation capture less intra-category variance and more inter-category variance (i.e., mostly between the object

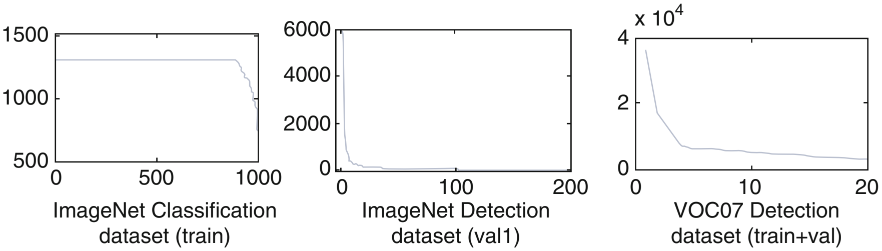


Fig. 3 The number of samples in *y*-axis sorted in decreasing order for different classes in *x*-axis on different datasets

category and background), causing many false positives between ambiguous object categories (e.g., classify horses as cows).

Long-tailed distributions of data have been studied in object detection [69], scene parsing [70], and zero-shot learning [71]. Ouyang et al. [69] investigated the factors that influence the performance in fine-tuning for object detection with long-tailed distributions of samples. Their analysis and empirical results indicate that classes with more samples will pose a great impact on feature learning. It is better to make the sample number more uniform across classes. In [70], much better super-pixel classification results were achieved by expanding the poor classes' samples. Bengio et al. [72] pointed out that poor classes can be beneficial for knowledge learned from semantically similar but richer classes. While in practice, other than learning the transfer features from richer classes, previous work mainly selects or simply replicates some of the data to avoid the potential long-tailed distribution problem. In [69], even if only 40% of positive samples are left out for feature learning, detection performance will improve slightly if the samples are uniform. The issue: To simply abandon part of the data, information contained in these identities may also be omitted. While some sampling heuristics may be applied, they are inefficient as the training procedure is still dominated by richer classes and there is room for further improvement.

5 Conclusion

Object detection is a key ability for most computers and robot vision systems. Although great progress has been observed in the last few years, we still notice that object detection has not been used much in many real-time applications where it could be of great help. Taking into account the speed of the object detection methods, while keeping the detection accuracy has gradually become the current research trend. Region-based methods have achieved good detection accuracy, but cannot satisfy the efficiency requirement in many practical applications. R-FCN is more computationally efficient than Faster R-CNN and well balanced in detection accuracy and speed. Although proposal-free methods (such as YOLO) can achieve real-time performance, detection accuracy is a concern when compared to region-based methods. SSD improves YOLO by taking into account both detection accuracy and real-time requirements. Finally, we need object detection systems for robots that will explore areas that have not been seen by humans, such as deep sea or other planets, and the detection systems will have to learn new object classes as and when they progressively encounter more objects. In such cases, a real-time open-world learning ability will be critical.

References

1. Haritaoglu, I., D. Harwood, and L.S. Davis, W/sup 4: real-time surveillance of people and their activities. IEEE Transactions on pattern analysis and machine intelligence, 2000. 22(8): p. 809–830.
2. Collins, R.T., A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, and P. Burt, A system for video surveillance and monitoring. VSAM final report, 2000: p. 1–68.
3. Geiger, A., P. Lenz, C. Stiller, and R. Urtasun, Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 2013. 32(11): p. 1231–1237.
4. Dollár, P., C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009. IEEE.
5. LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998. 86(11): p. 2278–2324.
6. Lowe, D.G. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on. 1999. Ieee.
7. Dalal, N. and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005. IEEE.
8. Roberts, L.G., Machine perception of three-dimensional solids. 1963, Massachusetts Institute of Technology.
9. Pellegrino, F.A., W. Vanzella, and V. Torre, Edge detection revisited. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2004. 34(3): p. 1500–1518.
10. Marr, D. and E. Hildreth, Theory of edge detection. Proceedings of the Royal Society of London B: Biological Sciences, 1980. 207(1167): p. 187–217.
11. Canny, J., A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, 1986(6): p. 679–698.
12. Rosten, E., R. Porter, and T. Drummond, Faster and better: A machine learning approach to corner detection. IEEE transactions on pattern analysis and machine intelligence, 2010. 32(1): p. 105–119.
13. Harris, C. and M. Stephens. A combined corner and edge detector. In Alvey vision conference. 1988. Manchester, UK.
14. Papageorgiou, C.P., M. Oren, and T. Poggio. A general framework for object detection. In Computer vision, 1998. sixth international conference on. 1998. IEEE.
15. Ojala, T., M. Pietikäinen, and D. Harwood, A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 1996. 29(1): p. 51–59.

16. Ojala, T., M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. 1994. IEEE.
17. Felzenszwalb, P.F., R.B. Girshick, D. McAllester, and D. Ramanan, Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 2010. 32(9): p. 1627–1645.
18. Yan, J., Z. Lei, D. Yi, and S.Z. Li. Multi-pedestrian detection in crowded scenes: A global view. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. 2012. IEEE.
19. Yang, Y. and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. 2011. IEEE.
20. Yan, J., Z. Lei, L. Wen, and S.Z. Li. The fastest deformable part model for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
21. Huang, K.-Q., W.-Q. Ren, and T. Tan, A review on image object classification and detection. Chinese Journal of Computers, 2014. 37(6): p. 1225–1240.
22. Everingham, M., L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman, The pascal visual object classes (voc) challenge. International journal of computer vision, 2010. 88(2): p. 303–338.
23. Hearst, M.A., S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, Support vector machines. IEEE Intelligent Systems and their applications, 1998. 13(4): p. 18–28.
24. Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 2012.
25. Deng, J., A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, Imagenet large scale visual recognition competition. (ILSVRC2012), 2012.
26. Uijlings, J.R., K.E. Van De Sande, T. Gevers, and A.W. Smeulders, Selective search for object recognition. International journal of computer vision, 2013. 104(2): p. 154–171.
27. Girshick, R., J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
28. Girshick, R. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision. 2015.
29. He, K., X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In European Conference on Computer Vision. 2014. Springer.
30. Ren, S., K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. 2015.
31. Dai, J., Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems. 2016.
32. Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
33. Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg. Ssd: Single shot multibox detector. In European conference on computer vision. 2016. Springer.
34. Changpinyo, S., W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
35. Wah, C., S. Branson, P. Welinder, P. Perona, and S. Belongie, The caltech-ucsd birds-200-2011 dataset. 2011.
36. Redmon, J. and A. Farhadi, YOLO9000: better, faster, stronger. arXiv preprint arXiv:1612.08242, 2016.
37. Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
38. He, K., X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

39. Huang, G., Z. Liu, K.Q. Weinberger, and L. van der Maaten, Densely connected convolutional networks. arXiv preprint arXiv:1608.06993, 2016.
40. Jégou, S., M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. 2017. IEEE.
41. Shen, Z., Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, DSOD: Learning Deeply Supervised Object Detectors from Scratch. arXiv preprint arXiv:1708.01241, 2017.
42. Takác, M., A.S. Bijral, P. Richtárik, and N. Srebro. Mini-Batch Primal and Dual Methods for SVMs. In ICML (3). 2013.
43. Zeiler, M.D. and R. Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision. 2014. Springer.
44. Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
45. Szegedy, C., S. Ioffe, V. Vanhoucke, and A.A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In AAAI. 2017.
46. Kim, K.-H., S. Hong, B. Roh, Y. Cheon, and M. Park, PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. arXiv preprint arXiv:1608.08021, 2016.
47. Huang, J., V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, and S. Guadarrama, Speed/accuracy trade-offs for modern convolutional object detectors. arXiv preprint arXiv:1611.10012, 2016.
48. Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár, Focal loss for dense object detection. arXiv preprint arXiv:1708.02002, 2017.
49. Sung, K.-K., Learning and example selection for object and pattern detection. 1996.
50. Rowley, H.A., S. Baluja, and T. Kanade, Neural network-based face detection. IEEE Transactions on pattern analysis and machine intelligence, 1998. 20(1): p. 23–38.
51. Dollár, P., Z. Tu, P. Perona, and S. Belongie, Integral channel features. 2009.
52. Shrivastava, A., A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
53. Simo-Serra, E., E. Trulls, L. Ferraz, I. Kokkinos, and F. Moreno-Noguer, Fracking deep convolutional image descriptors. arXiv preprint arXiv:1412.6537, 2014.
54. Loshchilov, I. and F. Hutter, Online batch selection for faster training of neural networks. arXiv preprint arXiv:1511.06343, 2015.
55. Wang, X. and A. Gupta. Unsupervised learning of visual representations using videos. In Proceedings of the IEEE International Conference on Computer Vision. 2015.
56. Torralba, A., Contextual priming for object detection. International journal of computer vision, 2003. 53(2): p. 169–191.
57. Sermanet, P., K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.
58. Szegedy, C., S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, Scalable, high-quality object detection. arXiv preprint arXiv:1412.1441, 2014.
59. Gidaris, S. and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In Proceedings of the IEEE International Conference on Computer Vision. 2015.
60. Shrivastava, A. and A. Gupta. Contextual priming and feedback for faster r-cnn. In European Conference on Computer Vision. 2016. Springer.
61. Bell, S., C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
62. Pinheiro, P.O., R. Collobert, and P. Dollár. Learning to segment object candidates. In Advances in Neural Information Processing Systems. 2015.

63. Shrivastava, A., R. Sukthankar, J. Malik, and A. Gupta, Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851, 2016.
64. Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision. 2014. Springer.
65. Cai, Z., Q. Fan, R.S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In European Conference on Computer Vision. 2016. Springer.
66. Najibi, M., M. Rastegari, and L.S. Davis. G-cnn: an iterative grid based object detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
67. Gidaris, S. and N. Komodakis. Locnet: Improving localization accuracy for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
68. Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009. IEEE.
69. Ouyang, W., X. Wang, C. Zhang, and X. Yang. Factors in finetuning deep model for object detection with long-tail distribution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
70. Yang, J., B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
71. Norouzi, M., T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G.S. Corrado, and J. Dean, Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650, 2013.
72. Bengio, S. The battle against the long tail. In Talk on Workshop on Big Data and Statistical Machine Learning. 2015.

Video Classification Methods: Multimodal Techniques

Amal Dandashi and Jihad Mohamad Alja'am

1 Introduction

In the past few years, events in the Arab world have been escalating rapidly. Citizen journalism has been rising, where citizens spontaneously document any event happening in their location, by utilizing smart phones to shoot videos and uploading them in raw form, online. Thus, Digital Arabic Content (DAC) production is at a new high, and specifically, Arabic videos are in raw form, unclassified, unannotated, and unused. There have been minor advances in developing systems dedicated to the classification and management of DAC. The predominant features in these systems include multimedia content exchange, open-and-share delivery platforms enabling digital audio-visual (AV) content between different users, semantic metadata retrieval and exploitation, and adaptive and personalized content discovery and delivery.

While technologies have been developed attempting to tackle the analysis of DAC, the majority of work aimed at classification of Arabic videos is based on textual annotation or closed caption text extraction and processing. The proposed system design consists of implementing multimodal video classification such that annotations and caption processing is excluded.

TRECVID (http://trecvid.nist.gov/) is a workshop series launched to promote and encourage research in information retrieval and video analysis by providing large test collections, forums, and uniform scoring procedures. Authors in studies [1–3] have proposed various video annotation tools that include automated and

A. Dandashi · J. M. Alja'am (✉)
Qatar University, Computer Science and Engineering Department, Doha, Qatar
e-mail: jaam@qu.edu.qa

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_3

semiautomated techniques, utilizing ontology languages, image-processing features and viewer comments, respectively. Khoury and others [4] have proposed the Semantic Video Content Annotation Tool (SVCAT) in an effort to address challenges in automated video annotation and usage of models that lack expressiveness. SVCAT is a semiautomatic annotation tool compliant with MPEG-7 standards. The novelty of the SVCAT lies in object localization via automated propagation, and metadata description via contour video tracking, thus alleviating the role of a human annotator.

Chu et al. [5] presented a semantic-based content abstraction and annotation method and a semantic pattern in an attempt to bridge the semantic gaps of content management. Sanchez et al. [6] introduced a methodology to partially annotate textual web content in an automatic and unsupervised way. It uses several well-established learning techniques and heuristics and relies on web information distribution. Jaoua et al. [7] developed a prototype of an Arabic search engine using formal concepts analysis (FCA) and Galois connection. Jaam et al. [8, 9] have developed new algorithms for Arabic text summarization, and automatic data classification. Elloumi et al. [10] developed data reduction and redundancy elimination algorithms, and knowledge extraction from Arabic and English news [11]. While technologies have been developed attempting to tackle the semantic analysis of digital content, there exists no global approach or solution to the low-level analysis of Arabic media.

Arabic digital data found on social media, information databases, and archives is vastly found uncategorized and in raw form. Large portions of this data are unused and irretrievable due to lack of content structuring. The lack of audio-visual search and retrieval tools for capturing unannotated digital media assets is the cause of the lack of reuse of this data. Most research done in the field of video classification and indexing is not language-oriented and does not accommodate the linguistic, social, and cultural specifications of Arabic digital media scenes. In this context, the objective of this study is to design, implement, and assess a multimodal system for automated classification of Arabic videos. The proposed system will target both raw and general-purpose Arabic content found on a variety of platforms on the web involving multimodal video processing. Raw material has some distinct constraints that must be considered during development: camera settings, shot-boundary detection, soundtrack irrelevance, redundancy detection, irrelevant annotations, and isolated fragments.

The proposed system will be composed of the following components: visual-based feature classification for event detection, Arabic named entity recognition (NER) tools for semantic audio content (dialogue transcription) classification, combined with audio-based analysis to extract patterns for event-based video classification. The classification domain targeted is "news" videos, pertaining to the events "shooting" and "explosion."

2 Background

Video classification is a wide and complicated field and consists of several modalities. The study in [12] is centered on segmentation techniques such as meanshift, graph-based, Nystrom, and segmentation by weighted aggregation (SWA), out of which the latter had the best result. In [13], the issue of vehicle detection and classification is tackled, by using multiple time spatial images obtained from a virtual detection line on a video frame.

Video segmentation and action recognition are usually viewed as independent problems [14]. This study proposes a method that avoids limitations of separating the two approaches by joint performance of video segmentation and action recognition. This method is based on a discriminative temporal extension of the spatial bag-of-words object-recognition-based model. The supervised methodology involves the invocation of a systematic and mathematically elegant time-series segmentation and action recognition algorithm. Classification is performed within a multiclass support vector machines (SVM) framework, while inference over the video segments is done with dynamic programming. Experimental evaluation was carried out on the Honeybee, Weizmann, and Hollywood datasets. While the proposed method depicted encouraging results on these standardized datasets, its reliance on fully labeled data poses a limitation to its usefulness to smaller training sets with few actions.

According to John and others [15], charting, a nonlinear dimensionality reduction algorithm, automatically estimates the intrinsic dimensionality of latent subspace, preserves local neighborhood and structure of high-dimensional data, and obtains both forward and inverse mapping. Charting is used in this study for articulated human motion classification in 3D data. This experiment involves classification of human action subsequences of varying lengths of skeletal poses, using a multilayered subspace classification scheme with layered pruning and searching. It also includes identification of minimum snippet length required for skeletal feature classification. Results depict similar or better classification accuracy than other comparable systems. However, authors note that increasing interclass discrimination would improve accuracy of shorter snippets and would be more efficient when the number of actions is high. This can be done by incorporating linear discriminant analysis within the charting framework.

In this paper [16], new results are presented in news video story segmentation and classification in the context of TRECVID video retrieval benchmarking event 2003. The Maximum Entropy statistical model is applied to fuse diverse features from multiple levels and modalities, including visual, audio, and text. Different features such as motion, face, music, speech, and high-level text segmentation information are included. Relevant features contributing to the detection of story boundaries are automatically discovered via the statistical fusion method. Using the large news video set from the TRECVID 2003 benchmark, results demonstrate a satisfactory performance has with the F1 measure up to 0.76.

3 Overview of Techniques

As the proposed approach involves the use of Arabic Named Entity Recognition to extract entities from speech, combined with audio-based pattern extraction for event detection, the background in this section will elaborate on the respective fields.

3.1 Named Entity Recognition

Named Entity Recognition (NER) was initially introduced as an information extraction technique. NER is a task that locates, extracts, and automatically classifies named entities into predefined classes in unstructured texts [17]. It covers proper names, temporal expressions, and numerical expressions. Proper names are classified into three main groups: persons, locations, and organizations. A class can be divided into subclasses to form an entire hierarchy, i.e., location can be classified into city, state, and country. The majority of NER studies have been focused on the English language, as it is the internationally dominant language, while research on other languages for the NER task has been limited.

Arabic is a richly morphological language of complex syntax. The lack of simplicity in the characteristics and specifications of the Arabic language makes it a challenging task for NER techniques. Arabic can be classified into three types: Classical Arabic, Modern Standard Arabic, and Colloquial Arabic. It is imperative for the task of NER to be able to distinguish between the three types. Classical Arabic is the formal version of Arabic used for over 1500 years in religious scripts. Modern Standard Arabic is that used in today's newspapers, magazines, books, etc. Colloquial Arabic is the spoken Arabic used by Arabs in their informal day-to-day speech and differs in dialect for each country and city. There are several specifications of the Arabic language that do not make NER an easy task: lack of capitalization, agglutination, optional short vowels, ambiguity inherent in named entities, and lack of uniformity in writing styles. Add to that common spelling mistakes and shortage of technological resources such as tagged corporas and gazetteers and we have several issues to tackle for tasks associated to natural language processing (NLP).

Many studies such as shown in [18–20] have tackled Named Entity Recognition, with the Arabic language domain as the center of the research. They have utilized different combination of features specific to the Arabic language, as well as different combinations of classifiers, such as the support vector machine, maximum entropy, and conditional random fields.

MADAMIRA is a natural language-processing system designed by Pasha et al. [21] that can be utilized for morphological analysis and disambiguation of Arabic text. It combines aspects of two previously used systems for NLP: MADA [22] and AMIRA [23]. MADAMIRA optimizes both previously mentioned systems with a more streamlined, robust, portable, extensible, and faster Java-based implementation. It includes several tasks useful for NLP processes: part-of-speech tagging, tokenized forms of words, diacritization, lemma stemming, base phrases, and NER.

3.2 Audio Analysis

Audio-only approaches [24] are more commonly utilized than text-only approaches for video classification. Audio approaches require fewer computational resources than that of visual methods. When features are stored, they also require less space. Another advantage is that segmented audio clips tend be very short (average 1–2 s), so the processing of the audio clips would be easier.

Audio features can lead to three layers of audio understanding: low-level acoustics, such as the average frequency for a frame; midlevel sound objects, such as the audio signature of the sound a ball makes while bouncing; and high-level scene classes, such as background music playing in certain types of video scenes.

Two main techniques use either time domain features or frequency domain features. Using time domain means plotting amplitude of a signal with respect to time, while frequency domain means plotting amplitude with respect to frequency, which pertains to the spectrum of signal.

The volume standard deviation and volume dynamic range measure may be utilized for time domain features, i.e., sports has a nearly constant level of noise. Different classes of sounds may be categorized by setting certain thresholds. The zero crossing rate (ZCR) is the number of signal amplitude sign changes per frame. A high ZCR indicates high frequency, i.e., speech has a higher ZCR variability than does music. Silence ratio is the proportion of a frame with amplitude values measured with respect to some threshold, i.e., news has a higher silence ratio than commercials, and speech has a higher silence ratio than music.

Frequency domain suggests an energy (signal) distribution across frequency components. The frequency centroid approximates brightness, and is the midpoint of the spectral energy distribution, i.e., brightness is lower in speech than in music. Bandwidth is the measure of the frequency range of a signal, i.e., speech has lower bandwidth than music. The lowest frequency in a sample is the fundamental frequency, which approximates pitch, and may be used to distinguish between speaker genders, or to identify parts of speech such as introduction of a new topic. A frame that is not silent and does not have a pitch represents noise.

3.3 Visual Analysis

Most approaches using visual features [1, 25] extract features on a per-frame basis. To name some definitions, a video is a collection of images or frames. All frames within a single camera action make up a shot. A scene is comprised of one or multiple shots that form a semantic unit. Often, using features that refer to cinematic principles (using shots, scenes, etc.) is utilized in visual-based approaches. This includes using light levels as indicators of genres, motion to measure action, and average shot-length to measure video pace. For example, an action movie is likely to have short average shot length and a specific light level. One disadvantage in utilizing visual-based features for video classification is the huge amount of potential data. This challenge may be addressed by using key frames to represent shots. Key frames are measured to be the most important frames extracted within a certain shot; this is done to avoid redundancy of data and results in less memory allocation needs. Rather than extracting thousands of frames from each video, to store in memory and process, utilize only the key frames, and increase frame-processing efficiency. Another technique to avoid having excessive data to store and process is to use dimensionality reduction techniques.

The overall process of a video indexing and retrieval framework is outlined as follows: (1) structure analysis: to detect shot boundaries, extract key frames, and segment scenes; (2) feature extraction from segmented video units (shots or scenes): these features include static features in key frames, object features, motion features, etc.; (3) video data mining using the extracted features; (4) video annotation: using extracted features and mined knowledge to build a semantic video index. The semantic index together with the high-dimensional index of video feature vectors constitutes the total index for video sequences that are stored in the database; (5) query: the video database is searched for the desired videos using the index and the video similarity measures; (6) video browsing and feedback: the videos found in response to a query are returned to the user to browse in the form of a video summary, and subsequent search results are optimized through relevance feedback.

3.4 Combination-Based Approach

Many studies incorporate the use of several combinations [26] of text and audio and visual features in order to complement each technique and overcome weaknesses of each. The main challenge of utilizing features from different modalities is knowing how and when to combine these features [27].

Qi et al. [28] use audio, visual, and textual features to classify news streams into genres of news stories. Audio and visual features are utilized to segment and group video shots into scenes. Text processing is used after detection of text through closed captions or scene text detection. Support vector machine classifier is used to classify the news stories.

Jasinschi and Louie [29] classify TV shows using audio, visual, and textual features. The audio features are used to classify six categories; noise, speech, music, speech and noise, speech and speech, and speech and music. Visual features are utilized to detect commercials. Textual features segment noncommercial parts of the TV program via annotations in closed captions. Finally, all audio categories are combined to classify the TV program as financial news or talk show.

Roach et al. [30] extend on their previous work, which consisted of classifying videos using audio features, to include using visual features. Adapted Gaussian Models for Image Classification (AGMM) is the classifier used for linear combination of the conditional probabilities of visual and audio features. The video classes studied are news, commercial, sports, cartoons and music videos.

Rasheed and Shah [31] utilize cinematic principles with accordance to audio and visual features to classify movies by analyzing the movie previews. Intersection of hue, saturation, and value (HSV) color histograms are used to segment previews into shots. Motion per preview is then calculated by using the ratio of moving pixels to total pixels per frame (visual disturbance), for each frame per preview. After visual disturbance is plotted against average shot length, a linear classifier is used to distinguish action and nonaction movies. Then audio energy variation analysis is used to categorize action movies into those with fire or explosions, or without. Light intensity thresholds are used to classify movies as comedy, drama, or horror. Horror movies have low levels of light intensity while comedies have the highest light levels, and dramas are in the middle.

4 Audio-Content-Based Classification Work

There are several studies that have attempted video or multimedia classification using only audio signals. These methods can be classified based on different types of detection, such as music detection, genre detection, scene detection, event detection, emotion detection, and others that are more specific, like violence or hazardous circumstance detection. Many studies utilize audio in accordance with visual-based classification: combination-based approaches. Others utilize deep learning techniques for audio-based detection. In this section, we present previous studies pertaining to audio-based event detection, genre detection, scene detection, and some multimodal combination-based detection studies.

4.1 Audio Event Detection

The amount of user-generated multimedia digital data on the Internet has increased exponentially over the past decade. Among the most popular multimedia sites, YouTube reported that 300 h of digital recordings are uploaded every minute [32]. As of March 2015, there are 70+ million hours of watch time on YouTube. There are other popular Internet sites that report similar statistics. The uploaded recordings are mostly unannotated, and descriptions are limited to high-level metadata, like author name or a brief title. Audio-based event detection is vital for extracting descriptions of multimedia recording and content analysis of digital audio.

The authors in [33] propose a system framework for learning acoustic event detectors using only weakly labeled data. The study involves a demonstration of the problem being formulated as a Multiple Instance Learning problem. A two-framework solution is then proposed for solving multiple-instance learning, one based on support vector machines (SVM) and the other on neural networks. The proposed approach leads to less time-consuming and less expensive process of the manual annotation of data, in order to facilitate fully supervised learning. The system is able to recognize events and provide temporal locations of the events in the recordings. Results show that events like clanking, scraping, and children's voices are easily detectable using SVM and neural network approaches, whereas events such as drums, hammering, and laughing are harder to detect using both of those methods.

Another study [34] deals with the detection of audio events derived from real-life recordings. The authors develop a technique for detecting signature audio events based on identifying patterns of occurrences of automatically learned atomic units of sound, named Acoustic Unit Descriptors (AUDs). Experimental results demonstrate that the proposed methodology works well for individual event detection as well as their boundaries in complex recordings.

In this work [35], the authors present an exemplar-based method for audio-based detection, based on nonnegative matrix factorization (NMF), which is only considered in the context of audio event detection. Events are modeled as linear combinations of dictionary atoms, and mixtures as linear combination of overlapping events. The weights of the activated atoms serve as direct evidence for the underlying event classes. This eliminates the need for error-prone source separation after conventional audio event detection. The proposed work offers three main contributions: modeling training data through exemplars, artificially increasing the amount of training data via linear time warping of the spectra at various rates, and explicit modeling of background events like noise. Results yielding promising outcomes on standardized datasets however indicated problems with either overfitting and/or development test mismatches.

In this study [36], the authors lay out how the bag-of-words model commonly used for text or visual-based classification has also been applied to audio-based classification; bag-of-audio words (BoAW). The proposed BoAW method extracts audio concepts in an unsupervised way. This gives it the advantage over other methods as it can be utilized easily for a new set of audio concepts in multimedia videos, without going through tedious manual annotation. Features are extracted from one-dimensional audio signals at fixed length intervals. These intervals may not capture the full acoustic variation that characterizes a specific sound. Experimental results depicted that certain representation decisions in the bag-of-visual-words algorithm such as L1-normalization are not optimal for audio representation. Results also varied in dependence on the acoustic variation of the video.

4.2 Audio Violence Event Detection

These audio detection systems centered on detection violence are conventionally developed in order to provide audio-based surveillance to public areas, in order to prevent or detect crime. The factor that researchers focus on is to minimize the false alarm rate, in order to avoid unnecessarily alarming responsible personnel.

This study [37] proposes a technique for automatic space monitoring based solely on the perceived audio data. The main objective of this study is to detect abnormal hazardous events in a noisy background environment. The authors focus on events where dangerous situations take place in a metro station, such as screams, explosions, and gunshots. The aim is to help warn authorized personnel to take precautions or take actions to prevent crime and property damage. In order to do this, the false alarm rate must be to a minimum. The approach utilized is based on a two-stage recognition schema that both utilize HMMs and GMMs to extract the approximate density function of the corresponding acoustic class. The feature set used was MFCC augment with a second group of parameters based on the MPEG-7 audio standard. Performance evaluation reports high detection rates in terms of false alarm and miss probability rates.

Another study [38] is centered on audio classification specifically for events concerning citizen security in urban environments. The various events studied are: explosion, broken glass, shot, shout, and others. The objective of this study is to build and test a system that performs audio-event detection for these specific danger situations, using MFCC features and HMM-based representation of acoustic data. The system is trained off a dataset of recordings developed by the authors. Performance results achieved promising results but could stand to use much optimization by tweaking the feature parameters.

In another study [39], authors present a method of violent shot detection in movies. They utilize audio and video modalities to classify, separately at first, and combine them at the end. For audio-based detection, a weakly supervised method is used to improve classification accuracy, to detect whether the movie is violent or nonviolent. Then they tackle detecting the violent event more specifically using visual-based classification to detect motion, flame, explosion, and blood-related events. Probabilistic Latent Semantic Analysis (PLSA) is utilized for this technique, and they test results on five movies, comparing the results of PLSA with the SVM classifier. The authors found enhanced results with their technique.

4.3 Audio Genre Detection

Lui et al. [40] used sample audio signals at a specific frequency, and after segmenting and subdividing into overlapping frames, utilized the following audio features: nonsilence ratio, volume standard deviation, volume dynamic range, pitch standard deviation, and others. Results depicted that the features with the highest discriminatory power are frequency centroid, frequency bandwidth, and energy ratio. Classification was then performed using one-class-one-network structure. The audio samples were then classed into commercial, basketball, football, news report, and weather forecast categories.

Roach and Mason [41] have utilized audio from video for the purpose of genre classification. They used Mel-frequency cepstral coefficients, which are coefficients derived from a cepstral representation of an audio clip. This approach was utilized due to its success with speech recognition. The authors find that best results are achieved with 10–12 coefficients. Classification is performed with the Gaussian mixture model due to its effectivity for speaker recognition. The genres studied are fast-moving sports, cartoons, news, commercials, and music.

Dinh et al. [42] use a Daubechies four wavelet to seven sub-bands of TV show audio clips. Wavelet transforms are useful for reducing dimensionality and have good energy compaction. The audio features used are sub-band energy, sub-band variance, zero crossing rate, as well as two customized features—centroid and bandwidth. Classifiers used are the C4.5 decision tree, K nearest neighbor, and support vector machine. Clips of different lengths not higher than 2 s were tested, and depicted no significant difference in performance. The genres tested were vocal music shows, news, commercials, cartoons, and motor racing sports.

Moncrief et al. [43] utilize audio-based cinematic principles to distinguish between horror and nonhorror films. Variations in energy intensity were used to detect sound energy levels, which in this study are associated with feelings of surprise, alarm, apprehension, surprise followed by alarm, and apprehension progression to climax. These four types of sound were found to be effective to distinguish horror movies and even to distinguish scenes within a horror movie.

4.4 Audio Scene Detection

Although the majority of studies surrounding acoustic classification have conventionally focused on music and speech signal processing, the challenge of acoustic environmental or scene detection has received more attention over the past few years. Recent work has focused more on nonstationary aspects of scenic sounds, and various new features centered on that have been proposed. In addition to that, sequential learning methods have been used to account for long-term variation of environmental sounds.

This study [44] presents a challenge on the detection and classification of acoustic scenes and events. The authors ran a scene classification challenge, and two event detection and classification challenges, namely, the office live (OL) and office synthetic (OS). The objective was to highlight areas that need improvement, to the research community concerned with audio-based scene detection. Results depicted that, in the case of scene classification, simple systems can do relatively well; however, complex systems can bring performance to the levels achieved by human listeners. The strongest performers chose a diverse set of features, used temporal information, and often used SVMs for classification.

In this study [45], a survey is conducted targeting acoustic scene detection studies. This work is centered on three main themes: basic environmental sound-processing methods, stationary techniques, and nonstationary techniques. Stationary techniques are dominated by spectral features, which are easy to compute but have limitations in the modeling on nonstationary sounds. Nonstationary techniques obtain features pertaining to the wavelet transform, the sparse representation, and the spectrogram. The latter two get the best results for nonstationary environmental sound detection. MFCC features are also utilized, often in combination with several other features to boost classification accuracy. Nonstationary methods give the best results, but are the most computationally expensive.

The authors also point out that each paper in this field presented its performance evaluations with their own datasets, due to a lack of standard datasets available for testing. This makes it difficult to conduct a fair quantitative comparison of different approaches.

5 Audio Open Source Code Libraries

There are several open source software projects dedicated to progressing the audio-processing/classification research community. The open source audio feature extraction toolbox, Yaafe [46], consists of audio-processing tools that use an audio-based combination of features to achieve a statistical learning model, in order for future events to be classified. The Yaafe toolbox includes several intermediate representations such as spectrum, envelope, and autocorrelation. It also includes options for temporal integration. There are several studies that have utilized the Yaafe toolset for audio-based component to classify videos and music [47, 48].

There are other audio-processing libraries built on the base of Yaafe features such as Essentia [49], which consists of audio-processing features and works with the GAIA plugin that complements the ESSENTIA studio with classifiers. The ESSENTIA project is an open source C++ library for audio analysis and audio-based music data retrieval. Its algorithm collection includes audio input/output functionalities, statistical characterization of data, digital signal-processing blocks. The features provided in ESSENTIA include spectral, temporal, tonal, and high-level music descriptors.

6 Audio Datasets

An audio classification system must be tested on standardized audio datasets in order to evaluate performance results, in comparison with similar studies. Standardized audio databases developed for testing are limited in number. We present here an overview on several developed audio datasets utilized for baseline testing.

The TUT Acoustic Scenes 2016 database [50] has been developed for environmental audio research. It consists of binaural recordings from 15 different acoustic environments, indoor and outdoor. TUT Sound Events 2016, a subset of this dataset, contains annotations for specific sound events. It consists of residential and home environments and is manually annotated. The authors provide a description of the database content, the recording and annotation procedure, along with a protocol for cross-validation and setup and performance results of acoustic scene classification and event detection, with the use of MFCC and GMMs. The TUT Acoustic Scenes 2016 dataset includes 15 acoustic scenes, namely, bus, beach, restaurant, city center, grocery store, home, library, forest path, car, metro station, train, tram, park, residential area, and home. All audio segments are 30 s long.

Acoustic scene classification pertains to the recognition of the audio environment, with applications in technology requiring environmental awareness. The environment may be specified in terms of physical or social context, for example, park, house, office, meeting, etc. Other databases for acoustic scene development include the DCASE 2013 [51] and the LITIS Rouen Audio Scene dataset [52].

The 2010 community-based Signal Separation Evaluation Campaign (SiSEC2010) includes an audio dataset developed specifically for the task of baseline testing for speech and music audio-based classification [53]. It contains seven speech and music datasets, including datasets recorded in noisy or dynamic environments, along with the SiSEC 2008 datasets. The authors provide a protocol for testing of five main tasks, and an evaluation guide using different objective performance criteria.

Another database that targets speech-processing studies and evaluations is the Open-Source Multi-Language Audio Database for Spoken Language Processing Applications [54], which contains speech passages from YouTube, specifically 300 passages in three languages; English, Mandarin, and Russian. The Multichannel audio database [55] consists of an acoustic dataset with audio segments in various acoustic environments designed in order to measure impulse response.

7 Proposed Approach

Several studies have addressed the issue of multimodal fusion for video classification [27]. Some chose to combine all the features into a single vector, while others trained classifiers for each modality, combined them at the end, and used another classifier to make a final decision. Identifying which modalities to use in combination, and the technique to combine them is an issue that largely depends on the domain being studied.

In this study, since the domain being tackled is in the Arabic news domain, specifically targeting the shooting and explosion categories, we choose to combine the audio-based approach with the textual-based approach to improve accuracy of results. As there are Arabic transcription tools, such as the IBM 2011 GALE Arabic speech transcription system presented in [56] that may successfully automatically transcribe Arabic videos in the news domain, we propose to extract the text from speech for the video classification purpose using this tool. We will utilize Named Entity Recognition (NER) tools, such as the MADAMIRA system presented in [21] to process and classify the text retrieved from transcription of speech in Arabic news videos, in order to retrieve elements vital to the classification data we need, namely, location, date, persons involved, events happening, such as the events we aim to classify; involving shootings or explosions.

Next, audio-based classification techniques will be utilized for event-based classification. The Yaafe studio consists of audio-processing tools that use audio-based combination of features to achieve a statistical learning model, in order for future events to be classified. The NER results, along with the audio-based results are to be combined with visual-based classification results, via multimedia fusion classifiers to achieve a final classification decision. The fusion classifiers include set weights for each modality and will be utilized with a preset threshold to classify data accurately. Results are to be compared with baseline results obtained from large-scale experiments such as TRECVid [57].

8 Video Dataset

The proposed system must be evaluated using baseline results achieved from similar systems that have utilized multimodal video classification techniques to classify Arabic news videos. However, most studies aimed at Arabic video classification have focused on classifying via textual techniques based on retrieving closed caption text or annotations, such as shown in [58–62] and many others.

The Activ video dataset developed by Zayene et al. [62] has been developed to include 80 videos consisting of more than 850,000 frames, from four different Arabic news channels. However, this dataset has been designed specifically to retrieve text from closed captions and assess the performance of text detection, tracking, and recognition systems. The data are accompanied by detailed annotations for each text box. While this dataset would be useful for textual-based video classification, for this study it may not serve as an optimal base for evaluation as the objective is to classify raw as well as broadcast news videos utilizing text-from-speech, and audio classification.

As TRECVid has also not released an Arabic news dataset since 2006, and neither has any other source, we propose two main solutions: (1) Design a new Arabic dataset consisting of Arabic news broadcast videos as well as raw news-related videos, to set as a new baseline evaluation tool for future classification attempts. (2) Adopt a multilingual approach as shown in Fig. 1, such that we test results using Arabic as well as English Named Entity Recognition for speech-to-text processing, combined with audio- and visual-based event classification techniques that are by default language-independent.

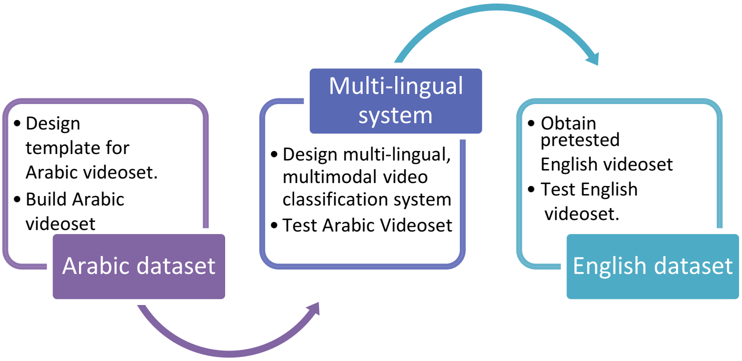

• Design template for Arabic videoset.
• Build Arabic videoset
Arabic dataset
Multi-lingual system
• Design multi-lingual, multimodal video classification system
• Test Arabic Videoset
• Obtain pretested English videoset
• Test English videoset.
English dataset

Fig. 1 Multilingual datasets

9 System Design

The proposed system design is composed of the following components, as illustrated in Fig. 2:

1. Dataset of videos obtained from Arabic news channel (i.e., AlJazeera) and social media (raw videos). This dataset is to be designed by the authors specifically to test multimodal video classification systems for the Arabic language. A different

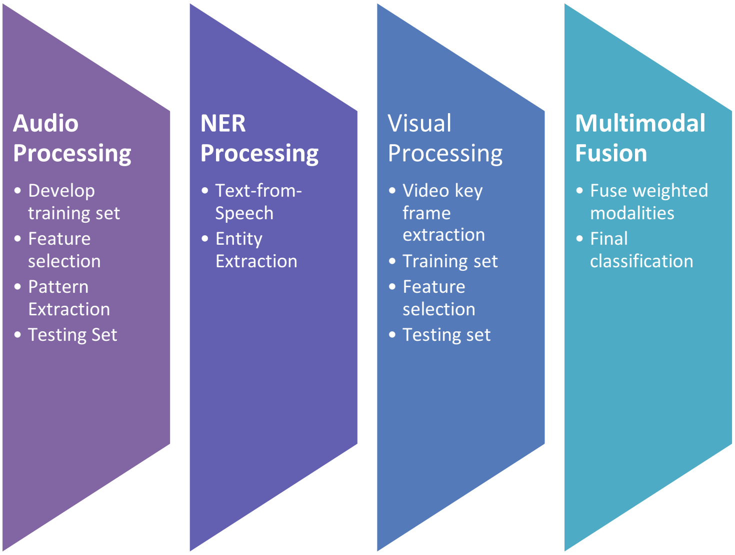


Fig. 2 System overview

dataset is to be used for the English video testing, from the TRECVid database. The domain to be detected is "news" with categories "explosion" or "shooting."

2. The NER-processing component consists of utilizing the NER component of the MADAMIRA system, to extract entities based on the speech-to-text conversion obtained from the videos. The entity categories to extract include: event, person, date, and location.
3. Audio-processing component: This step involves the utilization of audio-processing tools to classify videos based on frequency domain features. The open source audio feature extraction toolbox Yaafe [46] is to be utilized for audio classification pertaining to specific noises such as explosions or shootings. The Yaafe toolbox includes several intermediate representations such as spectrum, envelope, and autocorrelation. It also includes options for temporal integration.
4. Visual-processing component consists of utilizing visual-based combination of features for event detection. This step involved the extraction of key frames from videos, and utilization of visual-processing library to classify videos for event detection.
5. The multimodal fusion component is responsible for determining the correlation of the data, the confidence levels of the modalities, and the synchronization technique between different modalities. There are three main types of fusion: feature-level, decision-level, and hybrid multimodal fusion. Feature-level fusion refers to extracting features from multimodalities and combining them and sending them as input to a single analysis unit that performs the classification task. Decision-level fusion refers to features being processed individually and a local decision being made before combining them to use a decision fusion unit. The hybrid method consists of performing feature-level as well as decision-level fusion.

   There are several rule-based methods for multimodal fusion, the most commonly used being linear weighted fusion, specifically using the support vector machine classifier, as it can be easily used to prioritize different modalities. In this study, we will utilize the linear weighted fusion method to fuse the results of audio-based and NER-based classification.
6. The final step would be to perform exhaustive evaluation and testing, comparing the effect of using different combinations of features, as well as different classifiers. Testing will be compared on the Arabic dataset as well as the English dataset. Results are to be documented and graphed.

10 Conclusion

With the age of the Internet, increased usage of smart electronic devices and the exponential growth of social platforms for individual expression, digital multimedia has been increasingly uploaded in raw and unannotated form online. The need for technology systems to accurately classify and detect audio and multimedia has never been so vital. In addition to the need for classification, retrieval and reuse of user-uploaded audio clips, there are many other real-life applications that could stand to benefit from digital audio automated classification, some of those being music/movie platforms which could take advantage of audio-based genre classification, bank, surveillance and security applications that could benefit from acoustic scene or speech classification, hospital-based monitoring, military applications, and many others.

This study sheds light on information about the techniques and methodologies of audio classification, and presents a multitude of studies conducted with a focus on acoustic event, genre, scene, and combination-based classification. We also present a multimodal approach to classify raw videos, utilizing Arabic NER for processing text retrieved from speech, in order to extract entities related to persons involved, event, location, and date. A visual-based component is to be utilized for key frame-based event detection; an audio-processing component for extracting noise patterns among events like shootings or explosions is to be used in combination with the NER results. The final step consists of using a weighted multimodal fusion technique aimed to achieve optimal results.

Issues to consider in our future results pertaining to multimodal fusion include:

- Appropriate synchronization of different modalities
- Optimal weight assignment to different modalities
- Optimal integration of context into the fusion process
- Effective utilization of feature versus decision-level correlation
- Optimal modality selection

Acknowledgment This publication was made possible by GSRA grant # 1-1-1202-13026 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the author(s).

References

1. K. Khurana, and M.B. Chandak, Study of Various Video Annotation Techniques, In *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(1), 909–914, 2013.
2. D. Zhang, M.M. Islam, and G. Lu, A review on automatic image annotation techniques, In *Pattern Recognition*, *45*(1), 346–362, 2012.
3. P. Thompson, Viewer comments as educational annotation in video content sharing sites, In *International Journal of Social Media and Interactive Learning Environments*, *1*(2), 126–144, 2013.
4. V. El-Khoury, M. Jergler, D. Coquil, and H. Kosch, Semantic video content annotation at the object level, In *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia* (pp. 179–188). ACM. December 2012.
5. H. C. Chu, M. Y. Chen, and Y.M. Chen, A semantic-based approach to content abstraction and annotation for content management, In *Expert Systems with Applications*, *36*(2), 2360–2376, 2009.
6. D. Sánchez, D. Isern, and M. Millan, Content annotation for the semantic web: an automatic web based approach, In *Knowledge and Information Systems*, *27*(3), 393–418, 2011.

7. A. Jaoua, W. Labda, and J. Alja'am, Automatic Structuring of Arabic and English Search Engines Results Using Concept Analysis, In *International Journal of Computer Science and Engineering in Arabic. Vol. 3, No 01,* 2009.
8. J. ALJa'am, A. et al., Text Summarization Based on Conceptual Data Classification, In *International Journal of Information Technology and Web Engineering (IJITWE)*, *1*(4), 22–36, 2006.
9. A. Hasnah, A. Jaoua, and J. Jaam, Conceptual Data Classification: Application for Knowledge Extraction, In *Computer-Aided Intelligent Recognition Techniques and Applications*, 453–467, 2005.
10. S. Elloumi, J. Jaam, A. Hasnah, A. Jaoua, and I. Nafkha, A multi-level conceptual data reduction approach based on the Lukasiewicz implication, In *Information Sciences*, *163*(4), 253–262.2004.
11. S. Elloumi, et al., General learning approach for event extraction: Case of management change event, In *Journal of Information Science*, 0165551512464140, 2012.
12. Xu, C., & Corso, J. J. (2012, June). Evaluation of super-voxel methods for early video processing. *In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 1202–1209). IEEE.
13. Mithun, N. C., Rashid, N. U., & Rahman, S. M. (2012). Detection and classification of vehicles from video using multiple time-spatial images. *Intelligent Transportation Systems, IEEE Transactions on,* 13(3), 1215–1225.
14. Hoai, M., Lan, Z. Z., & De la Torre, F. (2011, June). Joint segmentation and classification of human actions in video. *In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 3265–3272). IEEE.
15. John, V., & Trucco, E. (2014). Charting-based subspace learning for video-based human action classification. *Machine vision and applications,* 25(1), 119–132.
16. Hsu, W., Kennedy, L., Huang, C. W., Chang, S. F., Lin, C. Y., & Iyengar, G. (2004, May). News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003. *In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (Vol. 3, pp. iii–645). IEEE.
17. D. Nadeau, and S. Sekine, A survey of named entity recognition and classification, In *Lingvisticae Investigationes*, *30*(1), 3–26, 2007.
18. Y. Benajiba, M. Diab, and P. Rosso, Arabic named entity recognition: A feature-driven study., In *Audio, Speech, and Language Processing, IEEE Transactions on*, *17*(5), 926–934, 2009.
19. I. Zitouni, X. Luo, and R. Florian, A cascaded approach to mention detection and chaining in Arabic, In *Audio, Speech, and Language Processing, IEEE Transactions on,* 17(5), 935–944, 2009.
20. I. Zitouni, and Y. Benajiba, Aligned-Parallel-Corpora Based Semi-Supervised Learning for Arabic Mention Detection, In *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(2), 314–324, 2014.
21. A. Pasha, et al., Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic, *In Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014.
22. N. Habash, et al., Morphological Analysis and Disambiguation for Dialectal Arabic. *In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, HLT-NAACL* (pp. 426–432), 2013.
23. M. Diab, Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. *In 2nd International Conference on Arabic Language Resources and Tools,* pp. 285–288, 2009.
24. M. H. Lee, S. Nepal, and U. Srinivasan, Edge-based semantic classification of sports video sequences, in *Proceedings of the International Conference on Multimedia and Expo,* vol. 2, pp. 157–160, 2003.
25. Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* 41(6), 797–819.

26. D. Brezeale, and D. J. Cook, Automatic video classification: A survey of the literature, In *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* 38(3), 416–430, 2008.
27. P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, In *Multimedia systems*, 16(6), 345–379, 2010
28. W. Qi, L. Gu, H. Jiang, X.-R. Chen, and H.-J. Zhang, Integrating visual, audio and text analysis for news video, In *Proceedings of the 7th IEEE International Conference on Image Processing (ICIP)*, pp. 520–523, September 2000.
29. R. S. Jasinschi and J. Louie, Automatic TV program genre classification based on audio patterns, In *Proceedings of the IEEE 27th Euromicro Conference*, pp. 370–375, 2001.
30. M. Roach, J. Mason, and L.-Q. Xu, Video genre verification using both acoustic and visual modes, In *International Workshop of Multimedia Signal Processing*, pp. 157–160, 2002.
31. Z. Rasheed and M. Shah, Movie genre classification by exploiting audiovisual features of previews, In the *IEEE International Conference of Pattern Recognition*, vol. 2, pp. 1086–1089, 2002.
32. Youtube statistics. http://www.youtube.com/yt/press/statistics.html.
33. A. Kumar, and R. Bhiksha. "Audio event detection using weakly labeled data." *In Proceedings of the 2016 ACM on Multimedia Conference*, pp. 1038–1047. ACM, 2016.
34. A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj. "Audio event detection from acoustic unit occurrence patterns." *In Acoustics, Speech and Signal Processing (ICASSP*), 2012 IEEE International Conference on, pp. 489–492. IEEE, 2012.
35. J. Gemmeke, L. Vuegen, P. Karsmakers, and B. Vanrumste. "An exemplar-based NMF approach to audio event detection." *In Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013 IEEE Workshop on, pp. 1–4. IEEE, 2013.
36. S. Pancoast, and M. Akbacak. "Bag-of-audio-words approach for multimedia event classification." *In Thirteenth Annual Conference of the International Speech Communication Association*. 2012.
37. S. Ntalampiras, I. Potamitis, and N. Fakotakis. "On acoustic surveillance of hazardous situations." *In Acoustics, Speech and Signal Processing,* 2009. ICASSP 2009. IEEE International Conference on, pp. 165–168. IEEE, 2009.
38. M. Pleva, E. Vozáriková, S. Ondáš, J. Juhár, and A. Čižmár. "Automatic detection of audio events indicating threats." *In IEEE International Conference on Multimedia Communications, Services and Security*, Krakow, vol. 6, no. 7.5. 2010.
39. J. Lin, and W. Wang. "Weakly-supervised violence detection in movies with audio and video based co-training." *Advances in Multimedia Information Processing-PCM* 2009 (2009): 930–935.
40. Z. Liu, J. Huang, and Y. Wang, Classification of TV programs based on audio information using hidden Markov model. In *Proceedings of the IEEE Multimedia Signal Processing Workshop,* pp. 27–32, 1998.
41. M. Roach and J. Mason, Classification of video genre using audio, In *Interspeech*, vol. 4, pp. 2693–2696, 2001.
42. J.-Y. Pan and C. Faloutsos, Videocube: A novel tool for video mining and classification, In *International Conference on Asian Digital Libraries*, pp. 194–205, Singapore, 2002.
43. S. Moncrieff, S. Venkatesh, and C. Dorai, Horror film genre typing and scene labeling via audio analysis, In *Proceedings of the International Conference on Multimedia and Expo,* vol. 1, pp. 193–196, 2003.
44. D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M.D. Plumbley. "Detection and classification of acoustic scenes and events: An IEEE AASP challenge." *In Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013 IEEE Workshop on, pp. 1–4. IEEE, 2013.
45. S. Chachada, and C-C. Jay Kuo. "Environmental sound recognition: A survey." *In Signal and Information Processing Association Annual Summit and Conference* (APSIPA), 2013 Asia-Pacific, pp. 1–9. IEEE, 2013.

46. B. Mathieu, et al., YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software, In *Proceedings of the 11th International Conf erence on Music Information Retrieval (ISMIR 2010).* 2010.
47. C. Frisson, et al., Videocycle: user-friendly navigation by similarity in video databases, In *Advances in Multimedia Modeling*. Springer Berlin Heidelberg, pp. 550–553. 2013.
48. C. Copeland, and S. Mehrotra, Musical Instrument Modeling and Classification.
49. D. Bogdanov, et al., ESSENTIA: an open-source library for sound and music analysis, In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013.
50. A. Mesaros, T. Heittola, and T. Virtanen. "TUT database for acoustic scene classification and sound event detection." *In Signal Processing Conference (EUSIPCO)*, 2016 24th European, pp. 1128–1132. IEEE, 2016.
51. D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.
52. A. Rakotomamonjy and G. Gasso, "Histogram of gradients of timefrequency representations for audio scene detection," *Tech. Rep., HAL*, 2014.
53. S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong. "The 2010 signal separation evaluation campaign (SiSEC2010): Audio source separation*." In International Conference on Latent Variable Analysis and Signal Separation*, pp. 114–122. Springer, Berlin, Heidelberg, 2010.
54. S. Zahorian. "Open-source multi-language audio database for spoken language processing applications." *STATE UNIV OF NEW YORK AT BINGHAMTON DEPT OF ELECTRICAL AND COMPUTER ENGINEERING*, 2012.
55. E. Hadad, F. Heese, P. Vary, and S. Gannot. "Multichannel audio database in various acoustic environments." *In Acoustic Signal Enhancement (IWAENC*), 2014 14th International Workshop on, pp. 313–317. IEEE, 2014.
56. L. Mangu, et al., The IBM 2011 GALE Arabic speech transcription system, In *Automatic Speech Recognition and Understanding (ASRU)*, 2011 pp. 272–277). IEEE, December 2011.
57. A. F. Smeaton, P. Over, and W. Kraaij, Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International workshop on Multimedia Information Retrieval* (pp. 321–330). ACM, October 2006.
58. M. Moradi, S. Mozaffari, and A. Orouji, Farsi/Arabic text extraction from video images by corner detection, In *Machine Vision and Image Processing (MVIP), 2010 6th Iranian*. IEEE, 2010.
59. M. Halima, H. Karray, and A. Alimi, A comprehensive method for Arabic video text detection, localization, extraction and recognition, In *Advances in Multimedia Information Processing-PCM 2010.* Springer Berlin Heidelberg, 648–659, 2010.
60. A. Anwar, G. Salama, and M. B. Abdelhalim, Video classification and retrieval using arabic closed caption, In *ICIT 2013 The 6th International Conference on Information Technology VIDEO*. 2013.
61. M. Halima, A. Alimi, and A. Vila, Nf-savo: Neuro-fuzzy System for Arabic Video OCR, In *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 10, pp. 128–136, 2012.
62. O. Zayene, et al., A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV, In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015.

Proposed Multi-label Image Classification Method Based on Gabor Filter

Ziad Abdallah, Ali El-Zaart, and Mohamad Oueidat

1 Introduction

Image Multi-label Classification (IMC) is an important topic in data mining that assigns a label or a set of labels to an image. The big demand for image annotation and archiving in the web attracts researchers to develop many algorithms for this application domain [1]. The Multi-Instance Multi-label Learning (MIML) is a framework of machine learning proposed recently for computer vision application [2]. In this framework, an image is described with many regions or instances and can be assigned to multiple labels.

For instance, Fig. 1 shows that the image contains three regions for the label "trees." Each region in the image is a set of instances. These regions can be expressed as different examples called feature vector, and in data mining it is called Multi-Instance Learning. At the same time, the image may be classified simultaneously for more than one label; it is then called Multi-Label Learning. The Multi-Label Learning models the relations between labels and regions (instead of the entire image).

This will decrease noises in this feature space and increase the accuracy of the model [1]. Figure 2 shows that only three regions in the image are assigned to three labels (sky, mountain, and water).

This comparison illustrated that multi-label learning takes into consideration the correlation between labels, while multi-instance learning connects regions to labels. MIML takes both relations simultaneously. MIML has been successfully applied to

Z. Abdallah (✉) · A. El-Zaart
Department of Mathematics and Computer Sciences, Beirut Arab University, Beirut, Lebanon
e-mail: z.abdallah@bau.edu.lb; elzaart@bau.edu.lb

M. Oueidat
Maintenance and Industrial Engineering Department, University Institute of Technology, Saida, Lebanon

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_4

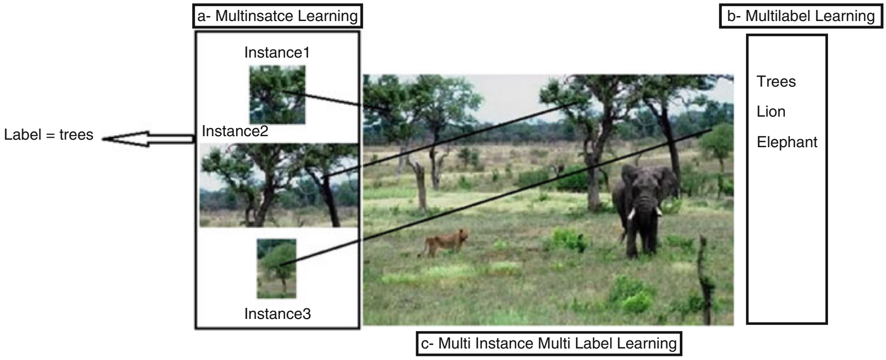


Fig. 1 Multi-Instance Multi-label Learning

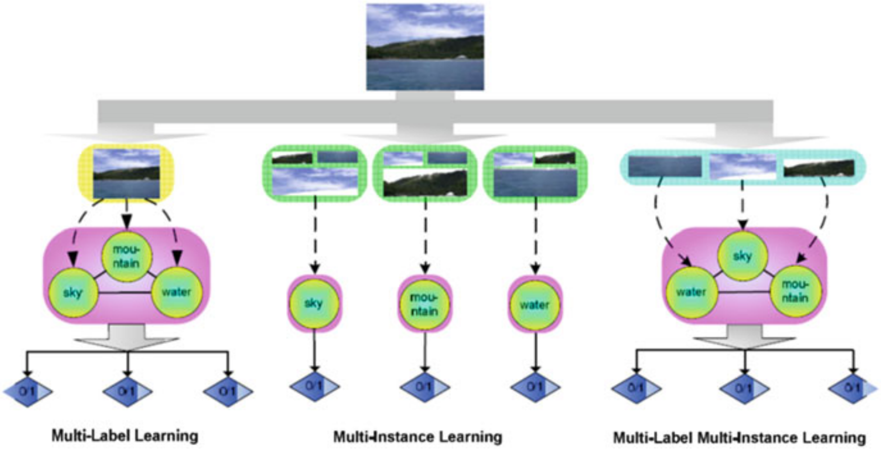


Fig. 2 Comparison between the three learnings [3]

image text classification, image annotation, video annotation, ecological protection, and other tasks [1–4].

The first way transforms multi-label to single-label. This transformation is called Multi-Instance Single-label Learning (MISL) and applies multi-instance learner to have Single-Instance Single-label Learning (SISL). The second way transforms multi-instance to single-instance. This transformation is called Single-Instance Multi-label Learning (SIML) and applies multi-label learner to have SISL. Two most important techniques are proposed for these transformations: MIML-Boost and MIML-SVM [4]. We are interested in this chapter in the second transformation as a multi-label problem. The drawbacks of these existing methods do not take into consideration the description of some characteristics from the image and the correlation between labels [5].

This chapter proposes a new framework that improves the MIML. The idea is to extract the feature from the image using Gabor filter bank (GFB). It is a feature extraction algorithm that takes into consideration the local representation, shape, and geometry of an image. Then we apply K-mean to cluster the image into similar groups. The final step consists of applying the Label Priority Power set as multi-label transformation in order to solve the problem of label correlation [6]. Each step of our new contribution is described in Sect. 3.

The remainder of this chapter is organized as follows: Sect. 2 defines the problem formulation. Section 3 defines the problem solution. Section 4 discusses the experimental results. Finally, the conclusions are presented in Sect. 5.

2 Problem Formulation

The MIML is an algorithm that transforms the problem to a single-label classification [4]. In Fig. 3, there are two ways to do this: the first one (T1) transforms MIML to MISL and applies Multi-Instance Learner (L1). The second one (T2) transforms MIML to SIML and applies Multi-Label Learner (L2). A good review can be found in [4]. Two most important techniques are proposed for these transformations:

1. MIML-BOOST: It is an MISL transformation. Each region in the image is transformed into a set of multi-instance called bags. Thus, original data-set can be split into a number of multi-instance datasets with only one label each. The learning task is transformed to traditional single-label learning. Figure 4 shows that the region in the original data-set is assigned to three labels (red green blue) from four (red green blue purple). In this algorithm, the region is split into four multi-instance regions.

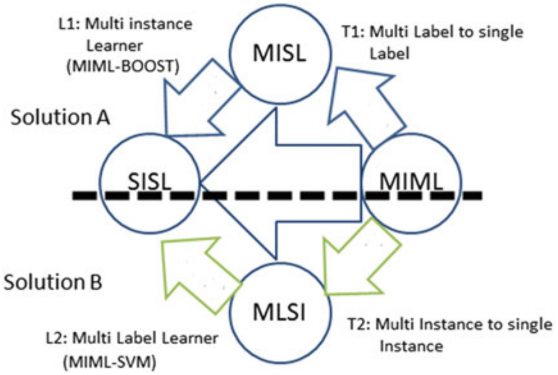

L1: Multi instance Learner (MIML-BOOST)
MISL
T1: Multi Label to single Label
Solution A
SISL
MIML
Solution B
MLSI
L2: Multi Label Learner (MIML-SVM)
T2: Multi Instance to single Instance

Fig. 3 MIML solutions [4]

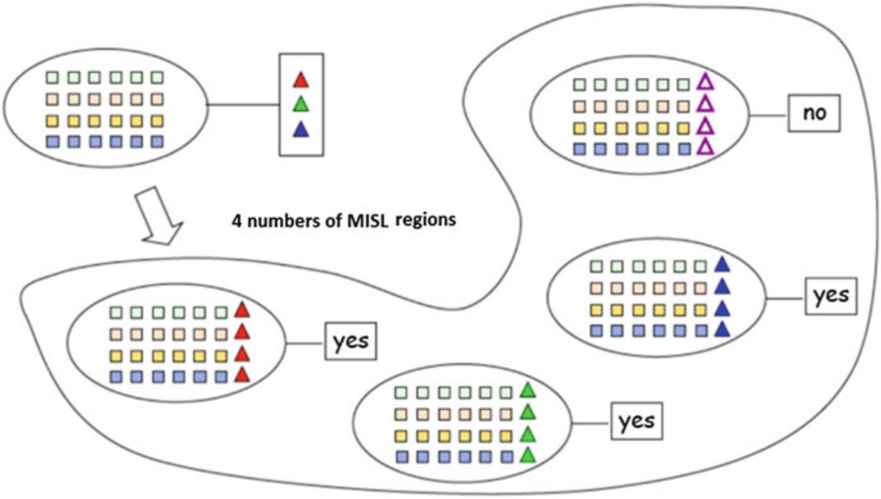


Fig. 4 MIML-boost illustrations [4]

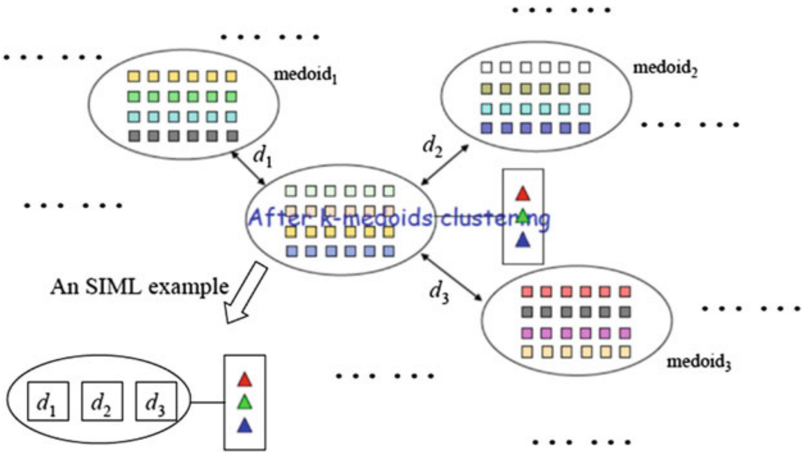


Fig. 5 MIML-SVM illustrations [4]

2. MIML-SVM: It is an SIML transformation. Each object is mapped into a feature vector using Hausdoff distance from a number of medoids generated first. The learning task is transformed to traditional multi-label learning.

Figure 5 shows that the region is mapped into a feature vector with three features computed using Hausdoff distance after using 3-medoids clustering algorithm. Figure 6 illustrates the steps of MIML-SVM.

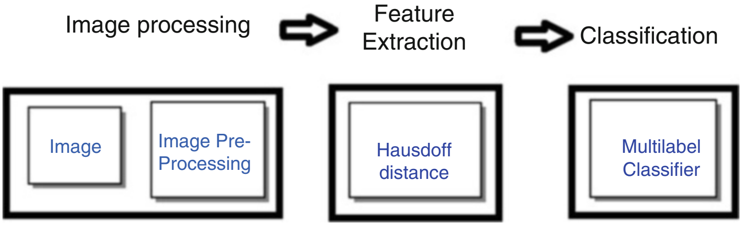


Fig. 6 MIML-SVM steps

The drawbacks of these existing methods are that they do not take into consideration the following:

1. The description of some characteristics from the image: color, shape, regions, textures and motion, and some elementary characteristics. The image could be assigned to sky or lake based on the relative size and placement of the components [5].
2. The correlation between labels: In multi-label learning, labels are correlated. For example, if the mountain label is assigned to the image with rocks and sky and the field label is assigned to image with grass and sky, then an image with grass, rocks, and sky would be assigned with both labels, field and mountain [5].

3 Problem Solution

3.1 Two-Dimensional Gabor Filter

This is a linear filter used in several domains in image processing [7, 8]. Formally, the two-dimensional Gabor filter family is expressed by the following equations:

$$g\left(x, y, \lambda, \theta, \psi, \sigma, \gamma\right) = \exp\left[-\frac{1}{2}\left(\frac{x'^2 + \gamma^2 y'^2}{\sigma^2}\right)\right] \exp\left[i\left(2\pi\frac{x^2}{\lambda^2} + \psi\right)\right] \quad (1)$$

$$x' = x\cos\theta + y\sin\theta \quad (2)$$

$$y' = y\cos\theta - x\sin\theta \quad (3)$$

Four parameters are important to determine the Gabor filter, as shown in Fig. 7, which shows the variation of

1. The wavelength of the sinusoidal factor λ
2. The orientation of the normal to the parallel stripes of a Gabor function θ

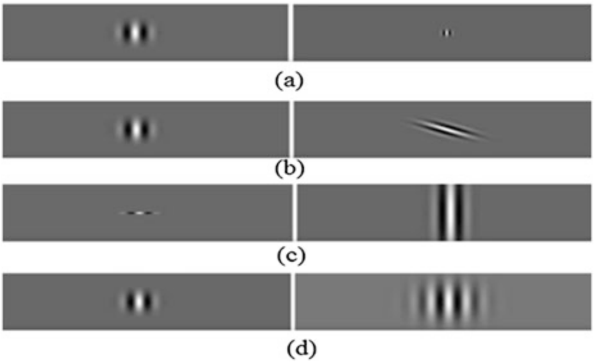
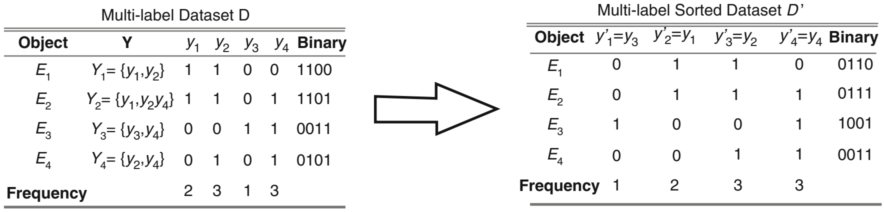


Fig. 7 Variations of the Gabor parameters [9]



Multi-label Dataset D

Object	Y	y_1	y_2	y_3	y_4	Binary
E_1	$Y_1 = \{y_1, y_2\}$	1	1	0	0	1100
E_2	$Y_2 = \{y_1, y_2 y_4\}$	1	1	0	1	1101
E_3	$Y_3 = \{y_3, y_4\}$	0	0	1	1	0011
E_4	$Y_4 = \{y_2, y_4\}$	0	1	0	1	0101
Frequency		2	3	1	3	

Multi-label Sorted Dataset *D'*

Object	$y'_1 = y_3$	$y'_2 = y_1$	$y'_3 = y_2$	$y'_4 = y_4$	Binary
E_1	0	1	1	0	0110
E_2	0	1	1	1	0111
E_3	1	0	0	1	1001
E_4	0	0	1	1	0011
Frequency	1	2	3	3	

Fig. 8 LPP transformation

3. The phase offset ψ
4. γ is the spatial aspect ratio

Moreover, σ is the sigma of the Gaussian envelope and usually it is equals one.

3.2 *The Label Priority Power Set (LPP)*

The LPP is a transformation method of multi-label learning into SISL [6]. It orders the label by importance. The advantage of this method is that it solves the problem of label correlation [6]. Figure 8 shows the conversion of multi-label data D to a multi-label dataset D' sorted by the frequency of each label.

3.3 *Framework MIML-LPPGABOR*

This section proposes a new framework that improves the MIML. The idea extracts important features (Mean, Standard Deviation, Skewness, Kurtosis, Entropy, First Quartile, Median, Third Quartile) from image using Gabor filter bank. It is a feature extraction algorithm that takes into consideration the local representation, shape, and geometry of an image. The first three features are the first central moments. They reflect the center position, the dispersion, and the asymmetry of the probability distribution. The drawback for these three features is that they are sensitive to outliers. Therefore, we add three features that divide the data in the image into four equal groups and they are not sensitive to outliers. Thus, we solve the first and second limitation of MIML. We then apply *K*-mean to cluster the image into similar groups. The final step consists of applying the Label Priority Power set as multi-label transformation in order to solve the problem of label correlation. The challenge of such learning is that the image contains many concepts existing in several regions at the same time. We faced the following issues:

1. The images do not have the same size [10].
2. Selection of the suitable feature from the image.
3. Different objects in the image could be similar [11].
4. Multiple objects in the same image.

Figure 9 shows block diagram of the new framework compared to MIML (Fig. 6). The following is a detailed representation:

Phase 1: Image Preprocessing
Resizing the image consists of changing the sample rate of the original image, preserving the important content and structure. Formally, let I be an image with m rows and n columns $I_{m\text{x}n}$. The resized image is an image $I'_{m'\text{x}n'}$. The output of this step is a dataset $D(I1_{m'\text{x}n'}, I2_{m'\text{x}n'}, \ldots, Ip_{m'\text{x}n'})$, where p is number of images in the dataset. All images in the dataset D have the same dimension. The advantage of this step is to prepare a dataset for the feature extraction process. The limitation is that only uniform scaling can be applied when resizing the image [10]. Figure 10 shows an example of resizing a sample of images.

Phase 2: Feature Extraction Using Gabor Filter Bank
Gabor filter bank (GFB) is composed of many distinct Gabor filters with different parameters. Two parameters are useful for extracting the suitable features from image [8]: the orientations and the frequencies. They are calculated using the

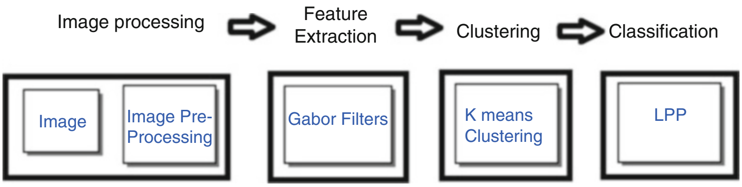


Fig. 9 Block diagram of the proposed method

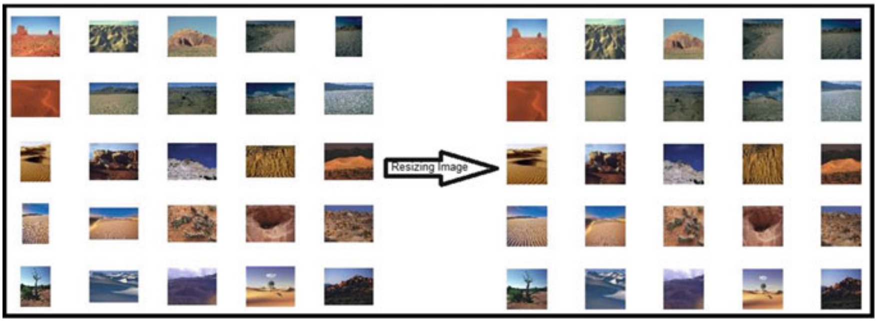


Fig. 10 Resizing a sample of images

following equations:

$$\theta(i) = \frac{(i-1)\,\pi}{\Theta}, \quad \text{where } i = 1, 2, \ldots \Theta \text{ (number of orientations)}. \tag{4}$$

$$\omega(i) = \frac{0.25}{\left(\sqrt{2}\right)^{\mathrm{i}-1}}, \quad \text{where } i = 1, 2, \ldots S \text{ (number of scales)} \tag{5}$$

Figure 11 shows GFB with five orientations and five frequencies.

The process of extracting feature from image, as shown in Fig. 12, consists of:

1. Reading the original image.
2. Resizing the image and transforming it from RGB to gray space color. The output is an image I with the size ($m = 128$, $n = 128$).
3. Applying each Gabor filter from GFB to I. Formally, this involves convolving each region in the image with the Gabor filter. The output of this step (c) is 25 filtered images with the same size as I.
4. Normalizing each filtered image by zero mean and unit variance. Then, it is subsampled by two factors: $d1$ and $d2 = (4,4)$, as in Fig. 12. That is, meaning that we will select 32 = 128/4 rows and 32 columns from the image. The output of this step is an image Is with the size (Ms = 32, Ns = 32). Each Is is partitioned into 4 × 4 Blocks. We extract nine features (Mean, Standard deviation, Skewness, Kurtosis, Entropy, First Quartile, Median, and Third Quartile) from each 2 × 2 blocks

Phase 3: *K*-Means Clustering

There are two kinds of cluster analysis techniques: *K*-Means and hierarchical Clustering. *K*-Means is better than hierarchical clustering in case of big amount of data [11]. *K*-Means consists of grouping similar images into different *k* mutually exclusive clusters. The output of this step is *K* clusters *C*1, *C*2, ..., *Ck*. An image

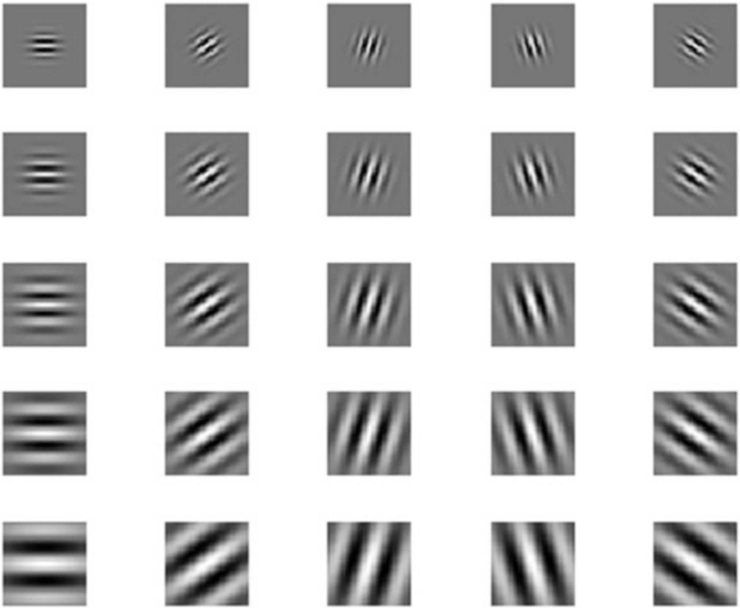


Fig. 11 GFB with five orientations and five frequencies

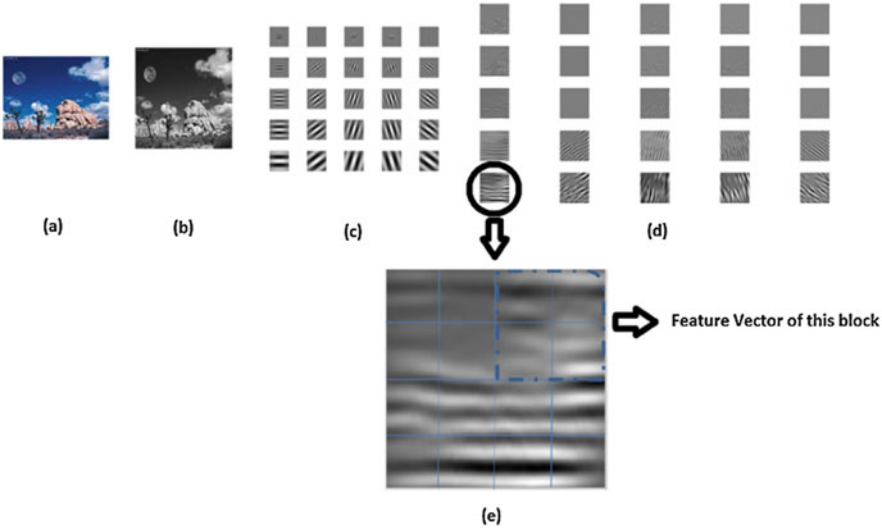


Fig. 12 Feature extraction using GFB

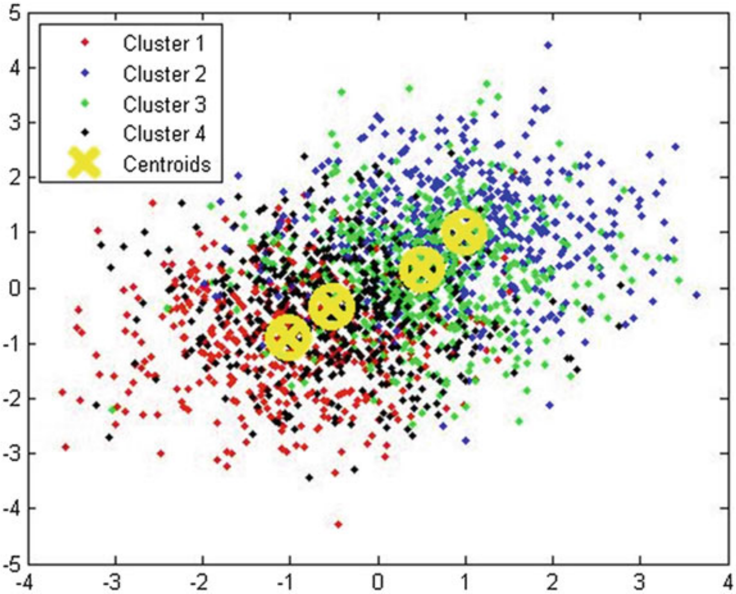


Fig. 13 Clustering method using K-Means

may belong to exactly one of these clusters. The advantage of *K*-Means is that it is better than hierarchical clustering in case of big amount of data when *K* is small. The disadvantage of *K*-Means is the difficulty of predicting the *K* representing the number of clusters. Figure 13 shows the centroids of four clusters generated by the 4-means algorithm.

Phase 4: Converting Multi-label Dataset to Multiclass Dataset using LPP
We will use in this step the transformation problem through breaking down the multi-label dataset into a single-label dataset using LPP transformation [6]. The output of this step is a dataset Ds = $\{(X1,y1), \ldots, (XP,yP)\}$, where *Xi* is the feature extracted from Gabor and *yi* is the decimal conversion of binary multi-label value as shown in Fig. 14. The importance of this step is the reduction of the complexity of learning process.

Tree Decision is a powerful classifier used in this phase because of its ease of use and its independence of the features of the dataset and their distribution. The output of this step is *k* trees, where *k* is the number of clusters. The advantage of this step is to applying single-label classification in a multi-label problem. Figure 15 shows the tree decisions constructed in the training phase for four cluster (the content of each tree is not important in the figure).

Multi-label Sorted Dataset *D*”

Object	Binary	yDec
E_1	0110	6
E_2	0111	7
E_3	1001	9
E_4	0011	3

Fig. 14 Conversions to decimal

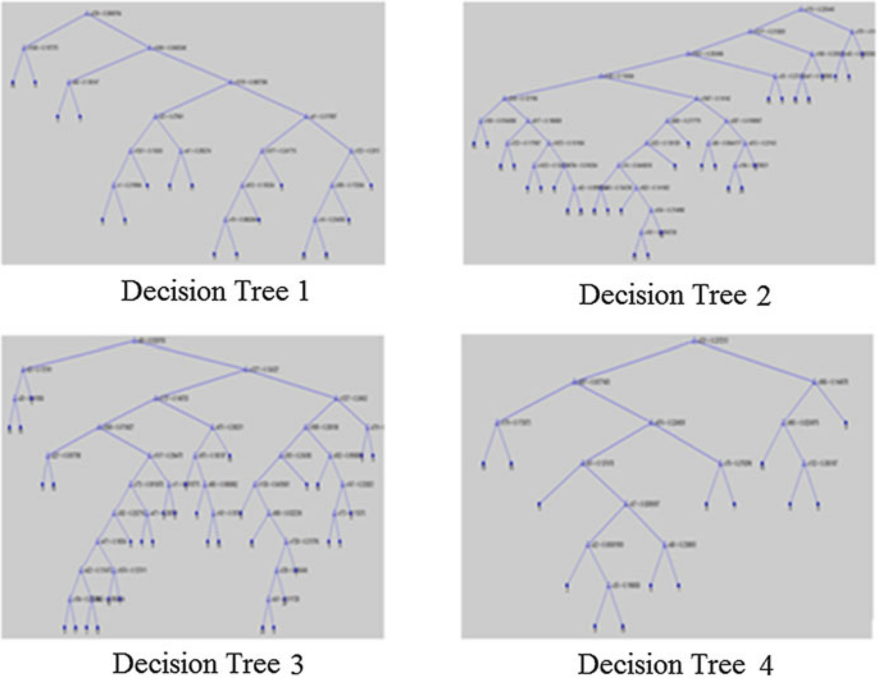


Fig. 15 Training phase

4 Experimental Results

Our contribution is built on the image multi-label classification domain. For this purpose, we use scene dataset. It is a benchmark used for this purpose for several state-of-the-art algorithms [4]. It consists of 2000 images belonging to five natural scenes: mountains, desert, sunset, trees, and sea. We split it into 1600 training examples and 400 testing examples. Therefore, five evaluation metrics are used: Hamming Loss (HL), Ranking Loss (RL), One Error (OE), Average Precision (AP) and Coverage [4, 12]. These metrics are commonly used to evaluate the performance of multi-label classification, taking into consideration:

- The mis-classification of examples–labels pairs (HL)
- The order of the proper label (RL and OE)
- Proper label ranked above particular label (AP)

It is clear from the above sections that there are many important parameters to be set up. We will discuss them in each phase.

Table 1 shows the values of parameters taken in the experiments. The size of the image 128 × 128 in the first phase and the couple (scale, orientation) is used in several references [8]. The parameter K should be small to have enough labels during the training phase.

The last parameter is the single-label classifier used in LPP. We used the decision tree as classifier. It is a powerful nonparametric method independent of the distribution of the feature vector space.

Table 2 presents:

- Better results compared to the five evaluation metrics (HL, RL, AP, OE, Coverage) according to the major MIML methods found in the literature. (MIMLBoost, MIMLSVMmi, and MIMLNN) [4].
- The results of our method using four parameters (Size of the image, Scale, Orientation, and the number of clusters K).

The analysis of this table shows a significant enhancement in all metrics using our method compared with the others of MIML. The transformation of multi-label to single-label gives better accuracy (single label metric) using LPP. This affects positively the results in all multi-label metrics.

Table 1 Parameters for each phase

Phase	Parameters	Values
Preprocessing phase	The size of the image	128 × 12864 × 64
Feature extraction	The scale, the orientation	(4,6) and (5,8)
K-Means	K	2, 3 and 4
LPP	Classifier	Tree decision

Table 2 Comparison between our method with the main state-of-the-art results

	HL↓	RL↓	AP↑	OE↓	Coverage↓
The best results (Algorithms[4])	0.185±0.08 (MIMLNN)	0.178±0.011 (MIMLBoost)	0.783±0.011 (MIMLSVMmi)	0.317±0.018 (MIMLSVMmi)	0.984.178 ±0.049 (MIMLBoost)
Size of the image is 128 × 128					
Scale_Orientation	K = 4				
(4, 6)	**0.0626**	**0.1212**	**0.9256**	0.0966	**0.5868**
(5, 8)	0.0727	0.1289	0.9222	**0.0905**	0.6379
	K = 3				
(5, 8)	**0.0589**	**0.1137**	**0.9325**	**0.0821**	**0.5693**
	K = 2				
(4, 6)	**0.0557**	**0.1089**	**0.9325**	**0.0795**	0.6073
(5, 8)	0.0615	0.1194	0.9272	0.0969	**0.5774**
Size of the image is 64 × 64					
Scale_Orientation	K = 4				
(4, 6)	**0.0606**	**0.1109**	**0.9323**	**0.0865**	**0.4861**
(5, 8)	0.0728	0.1341	0.9176	0.1118	0.5700
	K = 3				
(4, 6)	0.0726	0.1330	0.9187	0.1011	0.5612
(5, 8)	**0.0614**	**0.1114**	**0.9339**	**0.0867**	**0.5022**
	K = 2				
(4, 6)	**0.0604**	**0.1124**	**0.9337**	**0.0846**	**0.5090**

5 Conclusion

The aim of this chapter was to introduce a new framework for image multi-label classification. It is an improvement of the MIML framework. We presented the advantages of our method over three main MIML methods. The strengths of our method were found in its simplicity with regard to its implementation, solving the challenge of the description of the elementary characteristics from the image, and the correlation between labels and overall competitiveness in terms of the five evaluation metrics used. In the future, a new method can be developed in the feature extraction phase that optimizes the choice of each parameter.

References

1. R.S. Cabral, F. Torre, J.P. Costeira and A. Bernardino, Matrix completion for multi-label image classification. In Advances in Neural Information Processing Systems, pp. 190–198, 2011.
2. S.J. Huang., W. Gao and Z.H. Zhou, “Fast multi-instance multi-label learning”. In Twenty-Eighth AAAI Conference on Artificial Intelligence, 21 June 2014.
3. Z. J. Zha, X. S. Hua, T. Mei, J. Wang,,G. J. Qi and Z. Wang, “Joint multi-label multi-instance learning for image classification”. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1–8). IEEE, June 2008.
4. ML. Zhang and ZH. Zhou, “A review on multi-label learning algorithms.” IEEE transactions on knowledge and data engineering 26.8: pp. 1819–1837, August 2014.
5. M.R. Boutell, J. Luo, X. Shen, and C.M. Brown, “Learning multi-label scene classification”. Pattern recognition, 37(9), pp. 1757–1771, 2004.
6. Z. Abdallah, A. El-Zaart, and M. Oueidat. “An Improvement of Label PowerSet Method Based on Priority Label Transformation.” International Journal of Applied Engineering Research 11.16: pp. 9079–9087, 2016.
7. M. Nosrati, R. Karimi, M. Hariri and K. Malekian, “Edge detection techniques in processing digital images: investigation of canny algorithm and gabor method”. World Applied Programming, pp. 116–21, March 2013.
8. S. Khan, M. Hussain, H. Aboalsamh, H. Mathkour, G. Bebis and M. Zakariah, “Optimized Gabor features for mass classification in mammography”. Applied Soft Computing, pp. 267–80, 31 July 2016.
9. https://www.youtube.com/watch?v=-NZakhhB_Do
10. N. Sonawane and BD. Phulpagar, “Review on Content-Aware Image Re-sizing Using Improved Seam Carving and Frequency Domain Analysis”, 2015
11. M. Kaur and U. Kaur, “Comparison between k-means and hierarchical algorithm using query redirection”. International Journal of Advanced Research in Computer Science and Software Engineering, July 2013.
12. Abdallah, Ziad, Ali El Zaart, and Mohamad Oueidat. “Experimental analysis and comparison of multilabel problem transformation methods for multimedia domain.” Applied Research in Computer Science and Engineering (ICAR), 2015 International Conference on. IEEE, 2015

Vision-Based Approach for Real-Time Hand Detection and Gesture Recognition

Rayane El Sibai, Chady Abou Jaoude, and Jacques Demerjian

1 Introduction

Human-computer interaction (HCI) is the study of the interaction between humans and machines. It tries to improve the interaction between users and computers by making computers more usable and receptive to users' needs. Traditional devices for HCI such as keyboard and mouse are becoming feckless with virtual environment applications. So, an effective HCI with these environments requires more natural modalities. The most recent modality is the detection and recognition of hand gesture; it aims to build systems that can detect the user's hand and analyze it in order to control applications such as computer games, American Sign Languages (ASL) [1], TV controller [2], and 3D technology. However, hand gesture recognition is a challenging task; it contains various steps: image processing, hand detection and tracking, and gesture recognition. Several problems are facing these steps: (1) the presence of several persons in the scene other than the real subject, (2) the fact that the face of the user is seen by the camera, (3) the complex background, (4) the colors of the image which are sensitive to the lighting conditions, (5) non-white lights, and (6) color constancy.

The goal of this work is to develop a new approach for hand detection and gesture recognition by implementing a new algorithm that takes into account the problems

R. El Sibai (✉)
Université Pierre et Marie Curie, Paris, France
e-mail: rayane.el_sibai@etu.upmc.fr

C. A. Jaoude
TICKET Lab, Faculty of Engineering, Antonine University, Baabda, Lebanon
e-mail: chady.aboujaoude@ua.edu.lb

J. Demerjian
LARIFA-EDST Laboratory, Faculty of Sciences, Lebanese University, Fanar, Lebanon
e-mail: jacques.demerjian@ul.edu.lb

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_5

cited above. Hence, the system attempts to detect the user's hand(s) even if other people are present in the scene and even if his face is caught by the camera while trying to optimize the detection algorithm to be more robust to lighting conditions. After that, a set of features is extracted from the hand(s) detected and used to recognize the gesture performed.

The remainder of the chapter is organized as follows: Sect. 2 presents a review and a discussion of existing approaches for hand detection and segmentation. Section 3 exposes our proposed approach for real-time hand detection and gesture recognition. Section 4 discusses the testing phase of our prototype and the obtained results. The chapter ends with a conclusion and a description of future work.

2 Real-Time Hand Detection

In recent years, many systems based on the detection and recognition of hand gestures in real time have emerged. In desktop applications, hand gestures can be an alternative way of the mouse for the interaction between the user and the machine. Mouse gestures are used for various office tasks such as the manipulation of graphic objects, documents edition, web browsing, and so on. Using hand gestures for virtual reality applications becomes more and more recognized. Hand gestures allow the users to manipulate and interact in real time with 3D virtual objects using their hands. They are also involved in computer games. In such games, tracking the position of the hands or the body of the user can control the movement and orientation of interactive game objects.

Existing approaches for real-time hand detection can be classified into two categories: hardware-based approach and vision-based approach. When using the hardware-based approach, a contact device will be employed to detect and recognize the hand gestures. The devices employed to detect the hands and recognize the performed gestures are based on the physical interaction of the user with the interfacing device. These devices are generally based on technologies such as data glove, accelerometers, and multi-touch screens. CyberGlove is a wireless glove that uses various sensors to capture the motions of the hands and the fingers and transform them into data. In such an approach, there is often a motion tracker device attached to capture the position and rotation of data glove. This approach is not adaptable to naive users since it requires the user to be connected to the sensing device and to be habituated to use such a device. In addition, some health risks have been raised when using this approach: allergy symptoms caused by the devices such as mechanical sensors have been recorded [3].

A vision-based HCI system consists of three main phases: detection, tracking, and gesture recognition of the hand(s). The main challenge in HCI systems is the detection and isolation of the hand(s) of the user from each frame in the video. Following this segmentation, several features will be extracted from the isolated object in order to determine the gesture performed by the user. In this section, we discuss several methods that have been proposed in the literature for hand detection.

Hand detection based on skin color detection has been widely utilized in the literature [4–10] and [11]. For each input image captured by the camera, the skin color regions are detected in order to identify and to segment the hand(s) of the user. The main difficulty in such an approach is to determine the color space to be employed. Several color spaces can be used, including RGB, HSV, and YCbCr. The skin regions are detected according to specific predefined ranges of colors which depend on the color space used. A variant of this approach is to detect the hand(s) of the user based on the color of the glove that the user wears [12].

Color-based hand detection has several limitations. On the one hand, the segmentation of skin regions can be confused with other objects of the human body and with background objects having colors similar to that of human skin. One way to overcome this problem is to subtract the background before the detection of the hands. Hand detection based on background subtraction was introduced in [13] and [14]. The idea is to learn the scene's background and to subtract it from each incoming video frame. This subtraction is based on the assumption that the background is static and that the camera does not move. The background compensation methods can be applied to solve this problem [15]. On the one hand, the color of the human skin varies considerably across human races and even between humans of the same race, and it is affected by the lighting conditions and the characteristics of the camera. Thus, the correction of the colors and the brightness of the image are mandatory to compensate for this variability.

Sato et al. [16] proposed to use an infrared camera that detects the user's hand(s) based on a range of temperatures. The camera is installed with a surface mirror on the desk so that the hand can be observed by the camera. Then, the temperature's range of the camera is set to approximate the human body temperature (between 30° and 34°). Thus, the skin regions can be identified and extracted from the image using a threshold value. However, the whole human body has the same temperature; thus, if a body part is viewed by the camera, it will be also detected. Hand detection based on the shape is a popular method. It is independent of the skin color and the lighting conditions. With shape-based detection, several features are extracted from the contours of the objects in the image, thus isolating the user's hands based on several predefined characteristics [17]. Several approaches used the movement of the hands to detect them. They assume that the only movement in the image corresponds to the user's hands [18].

2.1 Discussion

In this section, we reviewed the existing methods for hand detection and segmentation. These methods are classified into two categories. When the hardware-based approach is used, the user needs to be connected to a sensing device and he must be habituated to use it. Vision-based methods can detect the hand more naturally and efficiently. However, when the hand detection is based on color classification, such as skin color and glove color, the detection result is very sensitive to the lighting conditions, background texture, and the existing objects in the scene. In the case of hand detection based on skin temperature, the detection result is very sensitive to the temperature of the existing objects in the scene. In the case of Haar-like features, a large set of hand images is required, and the process is time-consuming.

The robustness of a hand detection and gesture recognition system can be evaluated based on several measures. The adaptation of the system and its acceptability by the users is one of the main requirements of such systems. This implies the independence of the system of the user's type (age, skin color, clothing colors) and his experience with such systems. In addition, the gestures used by the system must be user-friendly, intuitive, and causing the slightest fatigue to the user.

The robustness of the system can be seen by its capacity to recognize the hand gestures effectively in different backgrounds. The background varies from a place to another depending on the conditions of the environment. These conditions change according to the illumination in the scene, the movements of the objects in the scene, and so on. As a conclusion, to achieve a precise and accurate hand detection result, three main conditions must be ensured: a good and sufficient brightness in the scene, color consistency of the video, and the ability of the system to differentiate between the user's hand(s) and other objects in the scene, such as the user's face and other people.

3 The Proposed Approach

Our proposed approach for real-time hand detection and gesture recognition is a vision-based approach. Our essential challenge is to develop a hand detection algorithm that satisfies the three conditions presented above. In our application, the frames are acquired through a webcam, and each input frame is processed to remove the unnecessary objects and highlight necessary components, which are the user's hand(s). After that, the hand(s) of the user are isolated, and a set of features is extracted and is used to recognize the performed gestures. Our system's architecture comprises two modules: hand detection module and gesture recognition module. The architecture of the system is shown in Fig. 1.

3.1 Hand Detection Module

Our hand detection algorithm is based on skin color detection. It distinguishes between the pixels having skin color and those which don't. After that, the module isolates the user's hand(s) by eliminating all other skin-colored objects detected. In order to acquire a better result and make the detection robust to the lighting conditions and colors' consistency, we perform two corrections for each frame in the video before segmenting the hand.

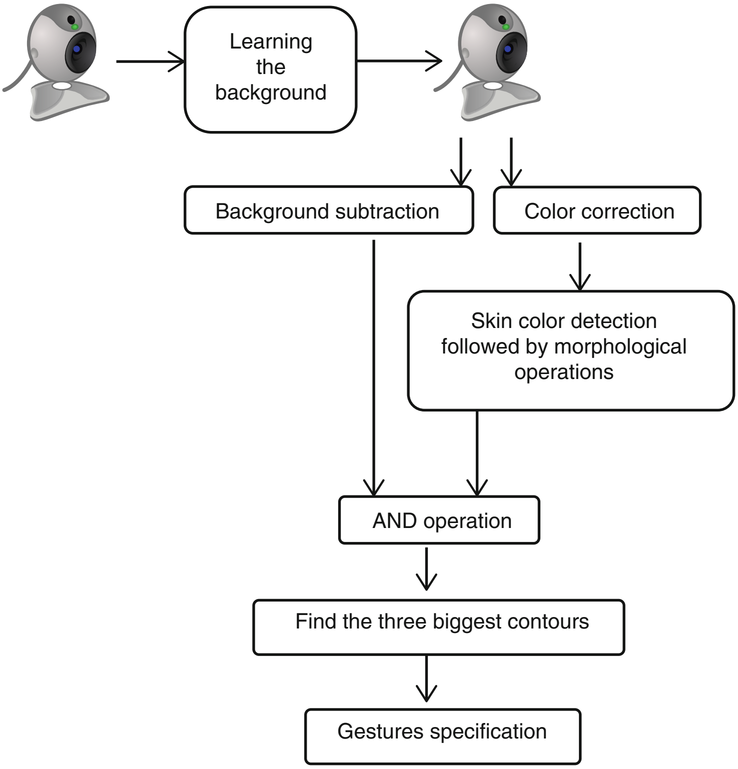


Fig. 1 Hand detection and gesture recognition system

3.1.1 Learning and Subtracting the Background

The first step in our application is to eliminate all unnecessary objects and isolate the user. The background must be eliminated to finally isolate the foreground which could be either a single user or multiple users. We suppose that the background is static, taking into account that the webcam should not move. Assume that there is a fixed webcam that captures a video in a room for a specific time, it will capture the tables, chairs, wall, etc. When a new person enters the room and is seen by the webcam, he is in the process of moving, which says that the person is a foreground, and other objects that are not supposed to move are defined as background.

To define the background, the user must launch the application without being seen by the webcam. During this time, the application saves a picture of all objects supposed to be immobile, and define it as a background. After that, the user can get to the webcam. The application proceeds to remove the learned background and keep only the new objects that appear in the scene. In case a new object other than the user is entering the scene, such as a chair, even if it is assumed stable, it will be considered as foreground because it was not seen and saved before by the webcam. In our application, the difference between each frame and the background is calculated in RGB color space. The result of this step is a binary image containing only the new objects entering the scene after saving the background.

3.1.2 Video Frame Color Correction

Specific color detection is a difficult step. Actually, the colors of an image recorded by a camera are sensitive to the lighting conditions, illumination level of the room, non-white lights, user position, and color consistency. They are also dependent on the physical objects present in the scene and the characteristics of the camera. These obstacles influence the colors of the recorded images and thus affect the performance of the skin color classification algorithm.

Therefore, it is necessary to correct the frame's colors before applying the skin detection algorithm. Color normalization of the image makes it possible to compensate for the discussed variations. This will make the detection less sensitive to the lighting conditions and improves the colors' consistency, which will ultimately make the skin color classification rate better. To achieve this goal, our developed Gamma correction function followed by the "Gray World" (GW) correction is applied consecutively:

- Gamma correction: The way to perceive the colors of an image varies according to the color and the intensity of the light. Thus, different materials respond to different lights in different ways. By correcting the $Gamma$, the image's brightness will be improved, making it appear brighter and more natural looking. Therefore, weak objects become more intense while objects of medium intensity weaken, unlike shiny objects. In order to have a dynamic $Gamma$ value according to the frame's brightness, we assume that each frame can have a brightness value between 16 and 235 and a Gamma value between 1.8 and 2.5. The proposed Gamma correction is described by Algorithm 1.
- Gray World (GW) correction [19]: Color consistency has been used in several domains for objects recognition such as robotics, bioinformatics, and artificial intelligence. The correction of the color consistency of the image makes it possible to recognize the true colors of the objects in the image and to correct them. The purpose of this adjustment is not only to correct the appearance of the image but also to improve the results of the object analysis, pattern and gesture recognition, medical imaging, and so on. GW is a simple, efficient, and widely used algorithm that aims to correct the image's colors and make them more consistent. It assumes that the average color of the image is gray and that the image has enough different colors, which is true for real-time images. It forces the frame to have a common average gray value for its red, green, and blue components.

Algorithm 1: Gamma correction: Image's brightness correction

```
Input  : Video frames
Output: Video frames with corrected brightness
foreach video frame do
    Calculate the brightness $B$;
    /* Calculate Gamma $\gamma$ from the average brightness $B$ */
    $\gamma = 2.3204 - (B * 0.00204)$;
    foreach pixel $i \in$ frame do
        /* $B'$ is the new frame's brightness */
        $B'_i = 255 * [\frac{B_i}{255}]^{\frac{1}{\gamma}}$;
    end
end
```

3.1.3 Skin Color Extraction

A wide range of colors can be created by subtracting the primary colors: cyan (C), magenta (M), yellow (Y), and black (K). The combination of several colors defines a specific color space. Since our detection algorithm is based on skin color, and since human skin color can be modeled by multiple color spaces, it is necessary to choose the best one. Once the frame's brightness and colors' consistency are corrected, each pixel will be classified as skin colored or not based on several predefined rules. These rules depend on the frame's color space. The output of this stage is a binary image, where each pixel i classified as skin region is represented by a white pixel and each pixel i classified as non-skin colored is represented by a black pixel. In order to get a better performance and choose the most efficient color space, three skin color detection algorithms based on RGB, YCbCr, and HSV color spaces were tested. More details about these masks can be found in [20].

3.1.4 Morphological and AND Operations

Binary images contain many imperfections. They are often distorted by noise and texture. The morphological treatment of the image is intended to eliminate these imperfections while taking into account the shape and the structure of the image. The output image of the previous stage is not clear; it needs some treatment to remove the noise and small contours. Three consecutive operations are applied: erosion to remove the fine structures, dilatation to fill the holes of the contours, and finally smoothing to eliminate the noise. Until now, there are two binary images: the first one contains the foreground subtracted, and the second one contains the skin-colored objects. AND operation between these two images is needed to obtain one binary image containing only the new skin-colored objects, as shown in Fig. 2.

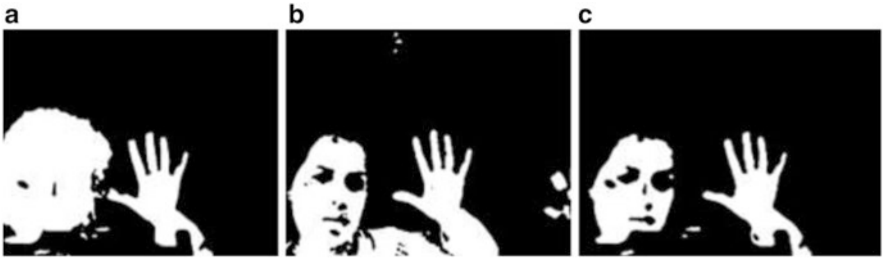


Fig. 2 (**a**) Foreground subtracted (**b**) RGB mask applied in the input image (**c**) AND operation

3.1.5 Find Three Biggest Contours

The result of the previous step is a binary image containing all foregrounds classified as skin colored. The next step is to isolate the objects corresponding to the user who is supposed to be the closest one to the webcam and eliminate all other objects. To do so, the contours of the object in the frame are detected. Once the contours are sought, their areas are calculated. The smallest ones are considered as noise and are deleted, and the three greatest ones are kept. As the user is the closest object to the camera, his hand(s) will always be among the three largest objects. In case the user is performing a one-hand gesture, the result is a binary image containing the three largest objects which are the user's hand and the other two biggest objects. In case the user is performing two-hand gestures, the binary image contains the user's hands and the other biggest object. Until now, the system isn't able to know if the user is performing one- or two-hand gestures.

3.1.6 Gesture Specification

This is the last step in the hand detection module, where the hand(s) of the user are isolated, and so, the system can specify if the user is performing a one- or two-hand gesture. By hypothesis, the user's hand(s) must always be lower than face level. If the user is performing two-hand gestures, his hands must be at the same level. Two parameters are used to isolate the hand(s) of the user. First, the application isolates the objects having the biggest value in the y-axis and then calculates their area and the distance between them. For this, two thresholds are defined: dy and dr.

3.2 Gesture Recognition Module

The goal of a hand detection and gesture recognition system is the interpretation of the hand location and posture, and to translate the gesture performed by the user into a specific action. Gestures can be static or dynamic. Dynamic gestures are a sequence of static gestures and they have a temporal aspect. Several approaches can be used to recognize the hand(s) gestures such as the hidden Markov models (HMM), supervised learning, and supervised learning with prediction [21].

In our work, the gesture recognition module is based on features extracted from the detected hand(s). The output image of the hand detection module is a binary image containing only the user's hand(s) performing the gesture. The hand(s) contour(s) should be approximated by another polygon having fewer vertices. This will reduce the number of unwanted convexity points which we later calculated. After the polygon approximation, the gesture is recognized by extracting and analyzing a set of features from the detected hand(s).

3.2.1 Features Extraction

Given the contour(s) of the hand(s), several features are extracted. The most important feature to use is the convex hull. It computes the hull of the detected contour and returns a list of convexities' defects assumed to be the hand's fingers. The shape of the object can be then characterized by a number of defect points. The area of each contour is approximated by a bounding box rectangle drawn around it; the position which defines where the object is located in the image along the vertical and horizontal axes. This feature is essential to track the hand and can be calculated from the center of bounding box rectangle drawn around the hand's contour. Finally, the rotation angle of the object can be extracted from the ellipse.

3.2.2 Gesture Recognition

Based on the features' values, gestures' recognition stage is done. If the output binary image of the hand detection module contains only one contour, this means that the user is performing a one-hand gesture. One-hand gestures are described in Table 1. If the output binary image of the hand detection module contains two contours, this means that the user is performing a two-hand gesture. Two-hand gestures are described in Table 2.

Table 1 One-hand gestures

Gesture	Hand's feature	Feature's variation
Open	Defects points	4–5 defects points detected
Move-to-right	x position	x value increases
Move-to-left	x position	x value decreases
Rotate-to-right	Angle	Angle value increases
Rotate-to-left	Angle	Angle value decreases

Table 2 Two-hand gestures

Gesture	Hand's feature	Feature's variation
Open	Defects points	8–10 defects points detected
Zoom-in	Difference of the x position between the two hands	x value increases
Zoom-out	Difference of the x position between the two hands	x value decreases

4 Experiments and Results

4.1 Software and Hardware Requirements

The system is developed using Visual Studio 2010 software. The framework OpenCV is used to develop the image processing, segmentation, and motion tracking functions. OpenCV is a free open-source framework developed by Intel and released under BSD license. It aims to develop real-time applications in computer vision domain. It is written in C, C++, Python, and Java interfaces, and it is supported by Windows, Linux, iPhone, and Android platforms. The essential functions provided by OpenCV are image processing and segmentation, camera calibration, and machine learning. It is an easy tool for experimenting with computer vision, and it is used in various applications such as gesture recognition, facial recognition system, motion tracking, object identification, and several HCI applications. The hardware requirements are minimal; the system consists of a laptop with a webcam.

4.2 Skin Color Detection

The objective is to test three skin color detection algorithms based on three color spaces: RGB, YCbCr, and HSV. Under different lighting conditions, these three algorithms are tested without the correction of the image's colors. The number of people in the scene is not important at this stage. When the skin color detection is based on the RGB color space, there is no need to do any conversion of the image. This is because the input image is given by default in the RGB model. Both YCbCr and HSV skin color detection algorithms need to convert the input image to the corresponding color space. The results show that regardless of the brightness of the scene, the skin color detection algorithm based on the HSV color space gives a very poor result. This is shown in Fig. 3. We have also noticed that the YCbCr and HSV skin color detection algorithms are slower than RGB mask. This is explained by the fact that they require a transformation of the input image from the given image color information to another one.

On the other hand, the results show that for very low brightness <70 and very high brightness >230, whatever the used color space, the result of the skin color detection is almost very poor. For a brightness between <70 and 190, RGB mask

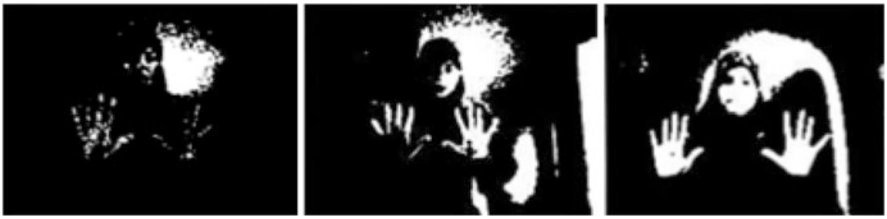


Fig. 3 Skin detection results based on HSV, YCbCr, and RGB color spaces, respectively

gives a much better result than YCbCr mask. For a brightness between 190 and 230, RGB and YCbCr masks give both very good results. The results also show that the RGB skin detection algorithm gives a sharp binary image. This is due to the efficiency of the algorithm with which we do not have to make any transformation from the current color space to another one because the input frame is given in RGB color space. Due to these results, our hand detection algorithm based on skin color detection will be based on the RGB color space.

4.3 Impact of the Colors' Correction on the Skin Detection Algorithm Based on RGB Color Space

The goal of color correction is to make the frame less sensitive to the illumination conditions and color inconsistency, and so, making the skin color detection algorithm more efficient and robust. Two corrections were applied to each frame. First, we applied our developed Gamma correction to adjust the brightness of the frame according to its current brightness. Then, we applied the "Gray World" correction in order to balance the colors of the image. Under different lighting conditions, RGB skin detection algorithm is tested on each frame before and after applying these corrections. The results have shown that for a very low brightness <70 and a very high brightness >230, the result of the detection remains very poor even after the correction. For a brightness between 70 and 190, the result will be better after doing the corrections, as shown in Fig. 4. For a brightness between 190 and 230, the brightness of the frame is already appropriate. In this case, if the frame's colors are also stable, the results of the skin detection (also called "mask") are very good with and without doing the corrections, because the distribution of colors and brightness of the image are already appropriate to having a good detection. However, if the colors of the image are not consistent, certain corrections are needed to rectify their consistency. After performing several experiments, the results showed that the optimal brightness in which the results are the best is between 90 and 115.

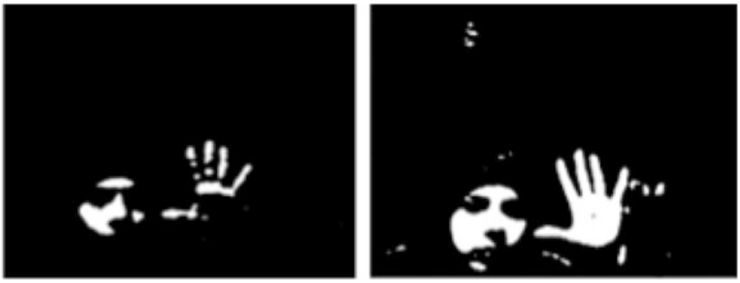


Fig. 4 RGB mask before and after the correction, under a brightness between 70 and 190

4.4 Hand(s) Segmentation and Gesture Recognition

In the scene, the face and the hand(s) of the user are always facing the webcam. In a more complex scenario, other faces and hands of other people may be also facing the webcam, so they will be detected and segmented by the skin detection algorithm. Therefore, the result will not only be the object of interest (ROI) which is the user's hand(s). Our algorithm is supposed to remove all unnecessary objects such as the faces, the hands of other people, and any other object detected by the skin detection algorithm. Our algorithm will then isolate the ROI and recognize the performed gesture by extracting and analyzing the features of the ROI. Under a brightness between 90 and 115, two scenarios were tested. For both scenarios, the corrections are applied for each frame, and the skin detection algorithm used is based on the RGB color space. The first scenario is a simple scenario, where only the user is facing the webcam. He is performing a one- or two-hand gesture, while the second scenario is a complex scenario in which many people are facing the webcam. The user is performing a one- or two-hand gesture. For the first scenario, the results show that our system is able to know if the user is doing a one-hand gesture or a two-hand gesture without any problem. For the second scenario, the results show that our application works very well, except when the user's hand(s) are overlapping with another hand(s) or face(s) in the scene. This requires that the user's hand(s) should not be confused with other skin objects facing the webcam.

4.5 2D Image Manipulation

After the hand(s) detection and segmentation and gesture recognition stages, the $2D$ object is manipulated. In our system, there are four gestures that allow the user to control a given $2D$ image. These gestures are "zoom-in" and "zoom-out" performed by two hands and "rotate-to-right" and "rotate-to-left" performed by one hand. Figures 5 and 6 show the results of a zoom-out gesture. The performed experiments show that the user is able to manipulate the image by performing the

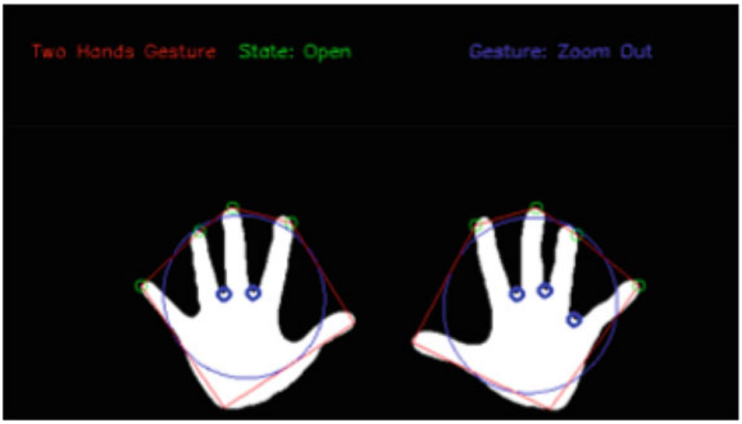


Fig. 5 Zoom-out gesture

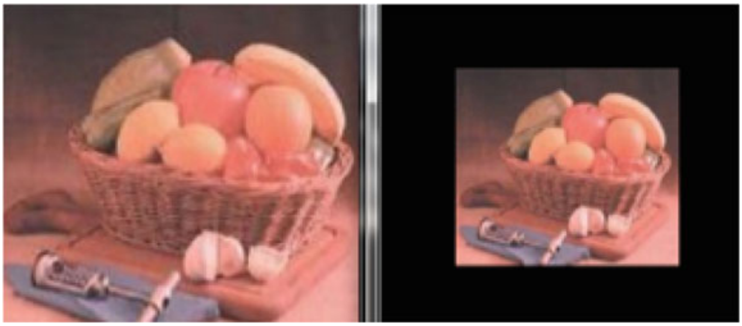


Fig. 6 Resizing image after applying a zoom-out gesture

specific gesture. However, in some cases, like the overlapping of hands, there is a false detection. The application segments an object which is not a hand, but may have 4–5 defects points, and classifies it as a hand.

5 Conclusion

In this work, a new approach for real-time hand detection and gesture recognition was introduced. The proposed approach can detect both hands of the user and make him able to interact with a 2D image, by performing the appropriate gesture. Our solution requires only a webcam. It tracks the hand(s) location in real time and recognizes several gestures. Our method can also detect the user's hand(s) even if his face or other people's faces are viewed by the webcam, and it can differentiate whether the user is performing one- or two-hand gestures. Our solution has overcome several limitations that existed in previous studies: the requirement that the user's face is not seen by the webcam, the prohibition of the presence of other people in the scene, and the need to wear a glove or the necessity of having the hand as the only object facing the camera. Also, our application can easily detect skin color regions even if the scene brightness is not adequate and even if the video's colors are inconstant. This was accomplished by applying the GW correction to correct the frames' colors and the Gamma correction to correct the video's brightness.

Nonetheless, our system can still be improved. We intend to ameliorate the hand(s) detection algorithm, optimize the background subtraction method to make it more robust whatever its color, and solve the problem of motion blur and hand trembling. Finally, we plan to make our system able to recognize more gestures, such as the click, grab, and swipe gestures.

References

1. Jayashree R Pansare, Shravan H Gawande, and Maya Ingle. Real-time static hand gesture recognition for American sign language (ASL) in complex background. *Journal of Signal and Information Processing*, 3(03):364, 2012.
2. Annamária R Várkonyi-Kóczy and Balázs Tusor. Human–computer interaction for smart environment applications using fuzzy hand posture and gesture models. *IEEE Transactions on Instrumentation and Measurement*, 60(5):1505–1514, 2011.
3. Maureen Schultz, Janet Gill, Sabiha Zubairi, Ruth Huber, and Fred Gordin. Bacterial contamination of computer keyboards in a teaching hospital. *Infection Control & Hospital Epidemiology*, 24(4):302–303, 2003.
4. Lawrence Y Deng, Jason C Hung, Huan-Chao Keh, Kun-Yi Lin, Yi-Jen Liu, Nan-Ching Huang, et al. Real-time hand gesture recognition by shape context based matching and cost matrix. *JNW*, 6(5):697–704, 2011.
5. Kui Liu and Nasser Kehtarnavaz. Real-time robust vision-based hand gesture recognition using stereo images. *Journal of Real-Time Image Processing*, 11(1):201–209, 2016.
6. TB Patil, Aakash Jain, Supriya C Sawant, Debnath Bhattacharyya, and Hye-Jin Kim. Virtual interactive hand gestures recognition system in real time environment. *International Journal of Database Theory and Application*, 9(7):39–50, 2016.
7. Hasup Lee, Yoshisuke Tateyama, and Tetsuro Ogi. Hand gesture recognition using blob detection for immersive projection display system. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 6(2):260–263, 2012.
8. Mokhtar M Hasan and Pramod K Mishra. Real time fingers and palm locating using dynamic circle templates. *International Journal of Computer Applications*, 41(6), 2012.
9. Hui-Shyong Yeo, Byung-Gook Lee, and Hyotaek Lim. Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware. *Multimedia Tools and Applications*, 74(8):2687–2715, 2015.
10. Manuj Paliwal, Gaurav Sharma, Dina Nath, Astitwa Rathore, Himanshu Mishra, and Soumik Mondal. A dynamic hand gesture recognition system for controlling vlc media player. In *Advances in Technology and Engineering (ICATE), 2013 International Conference on*, pages 1–4. IEEE, 2013.

11. Pushkar Dhawale, Masood Masoodian, and Bill Rogers. Bare-hand 3d gesture input to interactive systems. In *Proceedings of the 7th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: design centered HCI*, pages 25–32. ACM, 2006.
12. chandar Subash, Amalraj Willson, and sambandam Gnana. Real-time actuation of cylindrical manipulator model in opengl based on hand gestures recognized using open cvs. *International Journal of Modern Engineering Research*, pages 3497–3501, 2012.
13. Dong-Luong Dinh, Sungyoung Lee, and Tae-Seong Kim. Hand number gesture recognition using recognized hand parts in depth images. *Multimedia Tools and Applications*, 75(2):1333–1348, 2016.
14. Dariu M Gavrila and Larry S Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 73–80. IEEE, 1996.
15. Andrew Blake, Ben North, and Michael Isard. Learning multi-class dynamics. In *Advances in neural information processing systems*, pages 389–395, 1999.
16. Yoichi Sato, Yoshinori Kobayashi, and Hideki Koike. Fast tracking of hands and fingertips in infrared images for augmented desk interface. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 462–467. IEEE, 2000.
17. Erdem Yoruk, Ender Konukoglu, Bülent Sankur, and Jérôme Darbon. Shape-based hand recognition. *IEEE transactions on image processing*, 15(7):1803–1815, 2006.
18. Yuntao Cui and John J Weng. Hand sign recognition from intensity image sequences with complex backgrounds. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 259–264. IEEE, 1996.
19. Brian Funt, Kobus Barnard, and Lindsay Martin. Is machine colour constancy good enough? *Computer Vision—ECCV'98*, pages 445–459, 1998.
20. Mohamed Abdou Berbar. Novel colors correction approaches for natural scenes and skin detection techniques. *International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS*, 11(2):1–10, 2011.
21. Fayin Li and Harry Wechsler. Open set face recognition using transduction. *IEEE transactions on pattern analysis and machine intelligence*, 27(11):1686–1697, 2005.

Unsupervised Image Segmentation via Graph-Based Community Detection

Abdelmalik Moujahid, Fadi Dornaika, and Blanca Cases

1 Introduction

Scene parsing aims to segment and parse an image into different sub-images or regions associated with a variety of semantic objects. In aerial images, these objects are buildings, vegetation, roads, cars, and persons, among others.

There are two important questions in the context of scene parsing: how to produce good internal representations of the visual information and how to use contextual information to ensure the self-consistency of the interpretation [7, 21, 22]. These two issues are always addressed in a supervised context where a set of training images are used. In these images, the scenes are parsed and labeled manually. The supervision concerns two aspects: (1) the delineation of objects and (2) the categorization of the different objects found in the image.

This chapter presents an unsupervised method for scene clustering. The proposed framework is completely unsupervised in the sense that the scene partition is done without any training data. Our proposed method approaches the above two questions by transforming an over-segmented image into an optimal number of clusters that are more likely to be associated with semantic objects such as roads, vegetation, or buildings. The first step was to seek a suited image over-segmentation algorithm which is able to capture the contextual information needed for a group of pixels. Then, each of these groups has been described by a hybrid feature descriptor that

A. Moujahid (✉) · B. Cases
University of the Basque Country UPV/EHU, San Sebastian, Spain
e-mail: blanca.cases@ehu.es

F. Dornaika
University of the Basque Country UPV/EHU, San Sebastian, Spain

Ikerbasque Foundation, San Sebastian, Spain
e-mail: fadi.dornaika@ehu.es

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_6

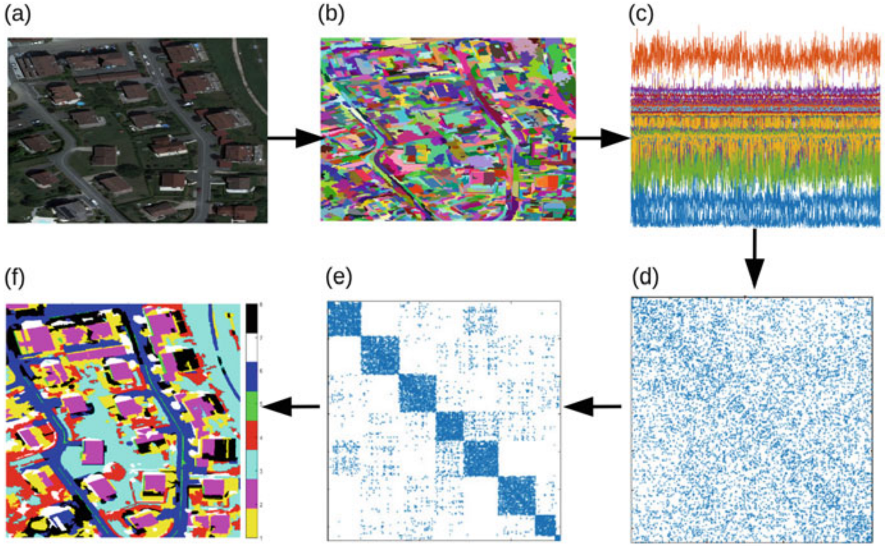


Fig. 1 (**a**) Original orthophoto. (**b**) Over-segmented image obtained using SRM algorithm. (**c**) The shape of the region descriptors. (**d**) The affinity matrix corresponding to the similarity graph. (**e**) Division of the graph into communities. (**f**) The resulting segmented image where the different communities have been associated with the different semantic objects in the original orthophoto

encodes rich information about color and texture. Finally, a community identification algorithm based on spectral modularity maximization criterion is performed on the descriptor graph [15]. This algorithm is completely unsupervised in the sense that the size and number of groups or clusters are unspecified in advance.

Figure 1 gives an overview of the proposed method. Starting from an original orthophoto (panel (a)), an over-segmentation is performed using the SRM algorithm resulting in 1591 homogeneous regions (panel (b)). Based on the proposed method, the number of these regions is clearly reduced to very few homogeneous regions or clusters that generally identify different categories in the image such as roads, roofs, or vegetation. The resulting segmented image is shown in panel (f) where the different communities have been associated with the different semantic objects in the image. Panels (d) and (e) show respectively the similarity graph before and after applying the community detection algorithm.

For the task of automatic scene parsing, the supervised approaches need training images in which a set of images are delineated and labeled by categories. Obviously, the manual labeling of only eight clusters is much more efficient and easier than labeling 1591 regions. This is one of the motivations of the proposed scheme.

The chapter is structured as follows. Graph construction methods are introduced in Sect. 2. Community detection algorithm and modularity definition are reported in Sect. 3. A detailed description of the proposed approach is reported in Sect. 4. The quantitative measures used to quantify the performance of the proposed approach are described in Sect. 5. Experimental results are described and discussed in Sect. 6. Finally, conclusions are drawn in Sect. 7.

2 Graph Construction Algorithms

In recent years, different graph construction methods have been reported, ranging from unified methods which calculate the graph and the embedding space in an iterative process to those that simultaneously seek a graph and a projection matrix (linear dimensionality reduction) [23, 31, 32].

In the following we briefly review two of the most popular methods for constructing a sparse graph from a set of examples or points: K-nearest neighbor (KNN) and locally linear embedding (LLE).

K-Nearest Neighbor Method involves two separate and independent processes: constructing the adjacency matrix from data and estimating the weights of the graph edges. For adjacency matrix construction, KNN can be used in order to find the neighbors of a datum. There is a function that defines the distance (similarity) of one input with respect to the others. In the second phase, a weight should be assigned to each constructed edge. In general, this weight should quantify the similarity between two connected nodes. Let $\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j)$ be the similarity score between neighbors $\mathbf{x}_i$ and $\mathbf{x}_j$; then the elements of the graph weight matrix $\mathbf{W}$ are given by Eq. (1). There are several choices for $\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j)$. For instance, [3] uses the heat kernel $\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}}$ with different Gaussian variance t values. In the extreme case where $t \to \infty$, the weights will become 0 and 1: 0 when there is no connection and 1 when two nodes are connected.

$$W_{ij} = \begin{cases} \mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

LLE Graph Construction Locally linear embedding (LLE) [18] formulates manifold learning problem as a neighborhood-preserving embedding, which learns the global structure by exploiting the local linear reconstructions. It estimates the reconstruction coefficients by minimizing the reconstruction error of the set of all local neighborhoods in the dataset. It turned out that the linear coding used by LLE can be used for computing the graph weight matrix. LLE graph can be obtained in two stages: adjacency matrix computation followed by the linear reconstruction of samples from their neighbors. The nonzero entries of the weight matrix $\mathbf{W}$ are estimated by reconstructing the sample from its neighboring points and minimizing the ℓ_2 reconstruction error defined as

$$\sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j} W_{ij}\, \mathbf{x}_j \right\|^2 \quad \text{s.t.} \sum_{j=1}^{n} W_{ij} = 1 \tag{2}$$

where $W_{ij} = 0$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are not neighbors.

Nonnegative LLE (NN-LLE) graph can be obtained by adding the non-negativity constraint on the coefficients W_{ij}.

3 Community Detection

Graph partitioning and community detection both refer to the division of the vertices of a graph into groups, clusters, or communities according to the natural pattern of edges in the graph. Most commonly one would to divide the vertices so that the groups formed are tightly linked with many edges inside groups and only a few edges between groups [2, 8, 20]. Some reviews on general clustering techniques can be found in [9, 11, 19, 29].

Graph partitioning is a classical problem in computer science [13] and consists in dividing the vertices of a graph into a given number of nonoverlapping groups of given size such that the number of edges between groups is minimized. Community detection, however, differs from graph partitioning in that the number and size of the groups into which the graph is divided are not specified in advance, which makes this method more suitable to solve real situations.

Over the past decade, spectral clustering attracted a lot of attention in the fields of data mining and pattern recognition [5, 10, 30]. It does not make assumption on the distribution of data, but directly estimates the global optimal solution on relaxed continuous domain through decomposition of the graph Laplacian matrix. Thus, it is simple to implement and can be solved efficiently and very often outperforms traditional clustering algorithms such as the K-means and K-medoids algorithms. Spectral clustering is an unsupervised technique. Some research works have introduced partial information into its framework by adding *cannot-link* and *must-link* constraints [26, 27].

In this work, we have adopted an algorithm for community detection based on spectral modularity maximization. Modularity is a measure of the extent to which like is connected to like in a network (or graph) and quantify how many edges lie within groups relative to the number of such edges expected on a random graph. Formally, modularity is computed as the difference between the actual and expected number of edges in the graph that join vertices of like type [14].

For a weighted graph **W** and a given partition of its nodes into C communities $\{c_1, c_2, \dots, c_K\}$, the modularity Q can be formulated as follows [14]:

$$Q = \frac{1}{2m} \sum_{ij} \left(W_{ij} - \frac{d_i d_j}{2m} \right) \delta(c_i, c_j) \tag{3}$$

where m is the volume of the graph **W** ($m = \sum_{ij} W_{ij}$) and $d_i = \sum_j W_{ij}$ is the total similarity between node i and all others. d_i is also called the degree of node i. $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise and c_i (c_j) is the cluster or community to which vertex i (j) belongs. This measure has high value when many edges in a graph fall between vertices of the same group than one would expect by chance. The main goal is to find good divisions of a graph into communities by optimizing Q over possible divisions.

The quantity

$$B_{ij} = W_{ij} - \frac{d_i d_j}{2m} \tag{4}$$

in Eq. (3) refers to the elements of a matrix **B**, which is called the *modularity matrix*, and it plays a role in the maximization of the modularity equivalent to that played by the Laplacian in standard spectral clustering. Unlike the Laplacian however, the eigenvalues of the modularity matrix are not necessarily all of one sign and in practice the matrix usually has both positive and negative eigenvalues. Let μ denote the number of positive eigenvalues (and corresponding eigenvectors) of **B**; then the maximum number of possible groups is given by $(\mu + 1)$ [15]. The eigenspectrum of the modularity matrix **B** is closely linked to the community structure of the graph.

Therefore, to reveal the community structure of the similarity graph, we proceed as follows: First, we retain the μ eigenvectors corresponding to the largest positive eigenvalues. Then, we iterate over $j = 1, \ldots, \mu$, spanning the whole range of possible groups. In each iteration, (1) we run a K-means algorithm on the retained eigenvectors looking for a partition into $C = j + 1$ communities, (2) we compute the corresponding modularity $Q(j)$ according to Eq. (3), and finally, we choose the optimal partition as the one with the maximum modularity $\max(Q)$.

Since the main goal is to maximize Q over possible divisions of the graph into communities, the expression of Q in Eq. (3) can be rewritten in terms of the $(n \times C)$ group membership matrix $\mathbf{S} = (\mathbf{s}_c)$ as follows:

$$Q = \frac{1}{2m} Tr(\mathbf{s}^T \mathbf{B}\mathbf{s}), \tag{5}$$

where Tr stands for the trace of a matrix, and

$$S_{ic} = \begin{cases} 1 & \text{if node } i \text{ is in community } c, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Alternatively, we have also analyzed the eigenspectrum of the traditional graph Laplacian matrices, but focusing on the set of eigenvectors corresponding to the smallest eigenvalues which contain the main information about the community structure in the graph.

We have considered both the unnormalized (**L**) and normalized ($\mathbf{L}_n$) matrices defined, respectively, as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \qquad \mathbf{L}_n = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} \tag{7}$$

where **W** refers to the similarity graph and **D** to the diagonal matrix of degree.

4 Graph Clustering Approach for Automatic Object Detection

The general flow diagram of the proposed method is illustrated in Fig. 2. The framework is composed of four main steps: image over-segmentation, feature descriptors extraction, graph construction, and community detection.

Over-segmentation is performed using the statistical region merging (SRM) algorithm [16], where the number of random variables (SRM parameter) has been fixed to $Q_{\mathrm{SRM}} = 2000$. This value has been chosen according to a comparative

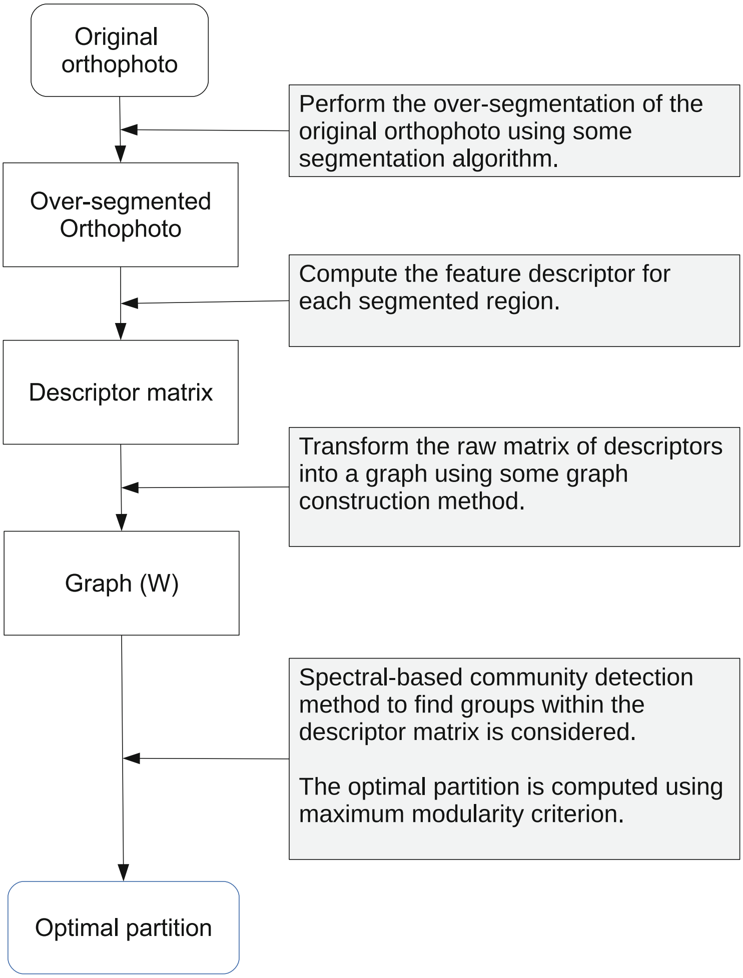

Original orthophoto
Perform the over-segmentation of the original orthophoto using some segmentation algorithm.
Over-segmented Orthophoto
Compute the feature descriptor for each segmented region.
Descriptor matrix
Transform the raw matrix of descriptors into a graph using some graph construction method.
Graph (W)
Spectral-based community detection method to find groups within the descriptor matrix is considered.
The optimal partition is computed using maximum modularity criterion.
Optimal partition

Fig. 2 Flow diagram of the full automatic detection of regions of interest (ROIs) in orthophotos. The algorithm broadly follows a four-step procedure

study we recently reported about new advances in image segmentation and region descriptors extraction for the automatic and accurate detection of buildings on aerial orthophotos [6].

Afterwards, each segmented region has been represented by a feature descriptor. It is well known that texture and color can characterize appearance variations of object surfaces. With this in mind, we used a hybrid descriptor that combines color histograms in RGB space and texture covariance descriptor.

To compute color histograms, we uniformly quantized each color channel into 16 bins and then the color histogram of each region is computed in the feature space of $16 \times 16 \times 16 = 4096$ bins.

The covariance descriptor, however, represents an image or an image region using sample covariance matrix [25]. Let J denote an $M \times N$ intensity or color image and V be the $M \times N \times d$ dimensional feature image extracted from J. Thus, V can be seen as a set of d 2D arrays (channels) where every array corresponds to a given image feature such as horizontal coordinate, vertical coordinate, color, image derivatives, and filter responses. This multidimensional array can be written as $V(x; y) = \phi(J; x; y)$, where ϕ is a function that extracts image features. Figure 3 shows a visual representation of this descriptor.

For a given image region $\mathscr{R} \in J$ containing n pixels, let $\{\mathbf{v}_i\}_{i=1\ldots n}$ denote the d-dimensional feature vectors obtained by ϕ within $\mathscr{R}$. According to [25], the region

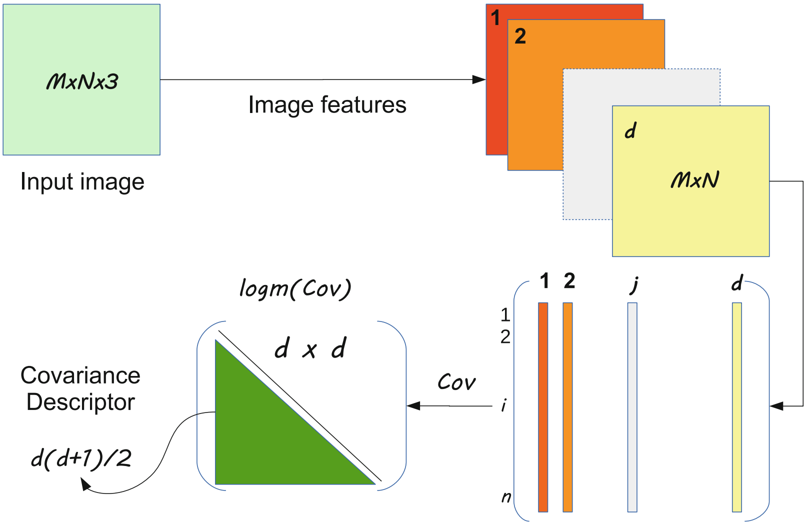


Fig. 3 A schematic representation of the covariance descriptor

$\mathscr{R}$ can be described by a $d \times d$ covariance matrix:

$$\Sigma_{\mathscr{R}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{v}_i - \bar{\mathbf{v}})(\mathbf{v}_i - \bar{\mathbf{v}})^T, \tag{8}$$

where $\bar{\mathbf{v}}$ is the mean vector of $\{\mathbf{v}_i\}_{i=1\ldots n}$.

Under the Log-Euclidean Riemannian metric, it is possible to measure the distance between covariance matrices. Given two covariance matrices Σ_1 and Σ_2, their distance is given by

$$d(\Sigma_1, \Sigma_2) = || \log(\Sigma_1) - \log(\Sigma_2)||_{\ell_2}, \tag{9}$$

where $||.||_{\ell_2}$ is the ℓ_2 vector norm and $\log(\Sigma)$ is the matrix logarithm of the square matrix Σ.

Thus, every image region, $\mathscr{R}$, can be characterized by $\log(\Sigma_{\mathscr{R}})$. Since this is a symmetric matrix, then the feature vector can be described by a $d \times (d+1)/2$ where d is the number of channels.

In our work, we consider 23 image features, that is, $d = 23$ [6]. Thus, the covariance descriptor is computed considering 23 channels including horizontal and vertical coordinates, 6 color channels both in RGB and HSV spaces, 6 image derivatives, and 9 local binary pattern (LBP) images [1, 17, 24] obtained by combining three different modes and three radii [4, 12, 28]. For all LBP images, the number of neighboring points is fixed to 8. Since the number of channels used is 23, it follows that the descriptor of each segmented region is described by $23 \times 24/2 = 276$ features. We stress the fact that even the covariance matrix descriptor uses color channels and LBP images, its descriptor is still different from that of color histograms and LBP histograms. A summary of the different types of descriptors and their dimensions is disclosed in Table 1.

Once all regions have been described by the concatenation of the color histogram and the covariance descriptor, a pairwise graph is built over all regions in the image.

Finally, based on the eigenspectrum of the resulting similarity graph, the over-segmented image is then partitioned into an unspecified number of similar clusters or groups using the community detection algorithm described in Sect. 3.

This approach aims to give the optimal partition of the feature descriptors datasets into communities or clusters such that the similarity between each pair of feature descriptors within each community is maximized and across communities is minimized.

Table 1 A summary of the different feature descriptors and their dimensions

Feature descriptor	Dimension
Color histogram	$16 \times 16 \times 16 = 4096$
Covariance	$23 \times 24/2 = 276$
Full descriptor	4372

To validate the proposed approach, we have also considered the traditional clustering algorithms such as K-means and K-medoids, which are applied directly to the region descriptors. In order to make these algorithms totally unsupervised, we have adopted the same criterion of maximum modularity to identify the optimal partition, which implies the use of the similarity graph obtained from data. For K-means and K-medoids clustering, the graph is used only to compute the modularity Q and does not intervene in the clustering process itself.

5 Performance Evaluation

The dataset used in this research to evaluate the performance of the proposed framework corresponds to 12 large orthophotos depicting several zones in the region of Belfort city situated on the northeastern of France. The spatial resolution of these orthophotos, provided by Communauté de l'Agglomération Belfortaine (CAB 2008), is (16 cm/pixel). These orthophotos contain about 200 buildings. In these orthophotos, the building roofs have different colors and textures. Furthermore, the background contains highly varying appearances corresponding to vegetation, cars, roads, and other objects. The sizes in pixels and the number of over-segmented regions in these orthophoto images are reported in Table 2.

Due to the fact that we know the ground-truth map of all building roofs in the orthophotos considered in this work, it is possible to quantify the agreement between the estimated cluster associated with buildings and the ground-truth data. For each orthophoto, we have computed the misclassification error (MCE) of the cluster that corresponds to the roofs and the entropy of the obtained partition by considering the obtained clusters and the two ground-truth classes: building and background. Note that the computation of the MCE was possible since we know the ground-truth of

Table 2 Orthophoto images size and number of regions obtained after over-segmentation

Orthophoto images	Size in pixels	Number of segmented regions
1	1712×1817	1536
2	1015×1735	1249
3	802×1297	1084
4	1796×1192	1273
5	738×884	900
6	2571×1503	1591
7	822×1094	1166
8	1780×1752	1551
9	1120×2516	1339
10	1254×1302	1315
11	963×1623	1181
12	1415×1489	1294

the roofs in all orthophotos. Thus, MCE is a supervised measure we have considered to quantify the performance of the algorithm in detecting roofs in the orthophotos and has been computed as

$$\mathrm{MCE} = \frac{\mathrm{card}(O \cup G - O \cap G)}{\mathrm{card}(O \cup G)}, \tag{10}$$

where O and G refer to the areas of the binary images representing the estimated optimal cluster (associated with roofs) and the ground-truth roofs, respectively. The computation of MCE requires identifying the roof cluster. This is automatically achieved by selecting the cluster that has the largest overlap with the ground-truth map of the buildings. A value of zero for MCE means that we have a perfect agreement between the estimated roof cluster and the ground-truth roofs.

Also, the entropy of a given partition has been computed based on the ground-truth maps as follows:

$$H = \sum_{i=1}^{N_c} \frac{n_i}{N} h_i, \tag{11}$$

where N_c is the number of identified clusters, N is the total number of individual regions, n_i is the number of regions in cluster i, and $h_i = -\sum_{j=1}^{2} p_{ij} \log_2 p_{ij}$ is the entropy of cluster i. $p_{ij} = \frac{n_{ij}}{n_i}$ refers to the probability that an element of cluster i belongs to class j, and n_{ij} is the number of elements in cluster i that are from class j where $j = 1$ denotes the building class and $j = 2$ denotes the background class.

In fact, in order to get a quantitative evaluation, we use the ground-truth building maps. The manually delineated buildings (ground-truth buildings) were used as a reference building set to evaluate the whole automated building-extraction accuracy. A zero entropy means that we have a perfect agreement between the automatic clustering of the two classes and the one provided by ground-truth data.

The performance of the proposed algorithm in detecting other objects in the orthophotos such as roads or vegetation has been evaluated qualitatively by inspecting the obtained clusters, as it can be seen in Fig. 4.

Likewise, we have also computed additional supervised measures such as F1 measure, accuracy, and Matthews correlation coefficient (MCC). The MCC returns a score between -1 and $+1$. A value of $+1$ means a perfect prediction, 0 no better than random prediction, and -1 means total disagreement between prediction and observation. The MCC is considered as being one of the best scores that can represent the confusion matrix of true and false positives and negatives by a single number. It should be noted that the recall and precision scores are also called "Completeness" and "Correctness," respectively.

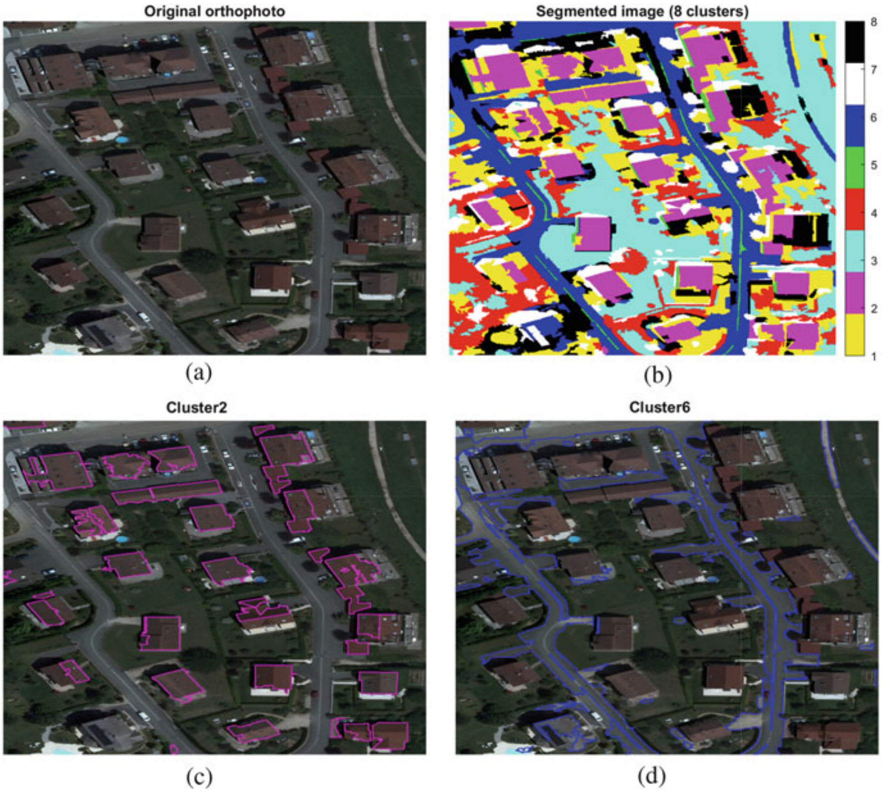


Fig. 4 (**a**) Original orthophoto. (**b**) The optimal partition into homogeneous regions obtained by the proposed clustering approach. The optimal partitions were selected according to the modularity maximum criterion. We can easily identify the different objects of interest in the images. Colors from yellow to black refer to clusters 1–8, respectively. (**c**) and (**d**) refer respectively to clusters representing mainly roofs and roads

6 Numerical Results

The segmentation process of the orthophoto regions, shown in Fig. 4a, is disclosed in panel Fig. 4b. The optimal clustering found has 8 groups or clusters. This partition has been achieved using a LLE similarity graph and adopting a spectral clustering based on modularity matrix **B**. As it can be appreciated from the obtained segmented image (panel (b)), clusters 1, 3, and 4 are mainly associated with vegetation, cluster 2 is associated with roofs, cluster 5 identifies the central line of roads and some dark area around buildings, cluster 6 mainly corresponds to roads, cluster 7 is mainly associated with the shadow of buildings and dark trees, and finally cluster 8 mainly corresponds to the areas around the houses including swimming pools, terraces, and cars. Although the method is totally unsupervised, the obtained clusters correspond to several semantic objects present in the scene.

Table 3 A quantitative evaluation of the semantic segmentation of the orthophoto shown in Fig. 4

	Ln	**L**	**B**	K-medoids	K-means
MCE					
Knn	0.27	0.27	0.27	0.41	0.60
LLE	0.25	0.26	0.26	0.43	0.44
NNLLE	0.25	0.26	**0.24**	0.47	0.44
Entropy					
Knn	0.18	0.18	0.17	0.26	0.31
LLE	0.17	0.15	0.18	0.27	0.26
NNLLE	0.16	0.18	**0.15**	0.28	0.27

Misclassification error (MCE) and entropy (associated with the building roofs) are reported. **Ln**, **L**, and **B** refer respectively to normalized Laplacian, unnormalized Laplacian, and modularity graphs. Minimum misclassification errors are shown in bold

A quantitative evaluation of that segmentation process for different graph construction methods and different spectral matrices is reported in Tables 3 and 5. Results obtained from these graph-based approaches have been compared with those achieved by the traditional K-medoids and K-means clustering algorithms. In this case, K-medoids and K-means have been performed on the descriptor data for different values of K, and the optimal partition has been chosen according to the maximum modularity criterion computed using the graph constructed from the descriptor data. For this reason, the results obtained by these algorithms have been also reported for different graph methods.

In order to quantify the quality of the obtained partition, we have computed quantitative measures such as modularity Q, misclassification error (MCE), and entropy (H) (see Table 3). The misclassification error (MCE) of the cluster corresponding to the roof and the entropy of this partition are also reported. The low values obtained for MCE and entropy indicate a good partition into homogeneous clusters. It is worthy to notice that best MCE and entropy were obtained with the nonnegative LLE similarity graph and the partition that is based on the modularity matrix **B**.

The qualitative analysis reveals that the obtained clusters are more likely to be associated with semantic objects in the image. Furthermore, the number of these identified clusters can be reduced if we merge modules with similar semantic information. For example, both clusters 3 and 4 correspond to vegetation and then can be merged into a single cluster (see Fig. 4a, b). To identify similar clusters we have computed the proximity matrix based on the average descriptors characterizing each cluster (see Table 4).

Table 4 Proximity matrix reporting distances between clusters for a partition of the orthophoto in Fig. 4a into eight clusters

Clust.	1 (y)	2 (m)	3 (c)	4 (r)	5 (g)	6 (b)	7 (w)	8 (k)
1 (y)	0	4.66	5.30	3.05	5.26	4.32	3.64	3.06
2 (m)	–	0	5.21	4.54	5.71	3.38	3.31	5.38
3 (c)	–	–	0	2.75	4.97	3.11	5.34	7.63
4 (r)	–	–	–	0	4.69	3.42	3.40	5.78
5 (g)	–	–	–	–	0	3.98	5.83	6.31
6 (b)	–	–	–	–	–	0	4.55	5.56
7 (w)	–	–	–	–	–	–	0	5.22
8 (k)	–	–	–	–	–	–	–	0

Table 5 Performance measures achieved by the algorithm in detecting the building roofs in the orthophoto shown in Fig. 4

	Ln	**L**	**B**	K-medoids	K-means
Accuracy					
Knn	95.24	95.24	95.24	92.63	88.94
LLE	95.56	95.39	95.31	91.87	91.70
NNLLE	95.52	95.37	**95.69**	89.55	91.21
F1 measures					
Knn	84.50	84.50	84.53	74.56	57.09
LLE	85.75	85.05	84.84	72.90	71.43
NNLLE	85.70	85.31	**86.22**	69.42	71.51
Matthews correlation coefficients (MCC)					
Knn	0.82	0.82	0.82	0.72	0.55
LLE	**0.84**4	0.83	0.83	0.69	0.68
NNLLE	**0.84**	0.83	**0.84**	0.63	0.67

Accuracy, F1 measure, and Matthews correlation coefficient (MCC) are reported. Minimum misclassification errors are shown in bold

On the other hand, using the ground-truth map of all building roofs in the orthophotos, we also have computed the following supervised measures: accuracy, F1 measure, and Matthews correlation coefficient (MCC) supervised measures. Results are depicted in Table 5.

Finally, to get a global validation of the proposed approach in detecting building roofs, we have computed the average over all orthophotos of the performance measures F1 and MCC (see Table 6). We can observe that the best results for F1 measure and MCC were obtained with the NNLLE graph and with partition that uses the spectral clustering based on the normalized Laplacian matrix. We can also observe that the performance obtained by the NNLLE graph and the modularity matrix **B** were also very good.

Table 6 Average performance measures in detecting the building roofs over all orthophotos

	Ln	**L**	**B**	K-medoids	K-means
F1 measures					
Knn	72.47	78.48	73.32	68.42	59.20
LLE	79.7	79.59	77.7	66.83	63.58
NNLLE	80.04	73.99	79.16	58.00	68.29
Matthews correlation coefficients (MCC)					
Knn	0.65	0.74	0.66	0.62	0.52
LLE	0.75	0.75	0.74	0.59	0.56
NNLLE	0.76	0.67	0.75	0.47	0.61

7 Conclusion

This work reports a new unsupervised approach for object detection in images by formulating the general problem of scene clustering as finding communities in networks or graphs. The proposed method has been tested both quantitatively and qualitatively on different orthophotos with promising results. Regarding building roof detection the proposed approach provides results that are in nice agreement with those achieved by a supervised approach reported recently in [6]. This indicates that the obtained clustering has good quality in terms of semantic labeling and geometric delineation.

Conflict of Interest Statement The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.
2. S. Bandyopadhyay, G. Chowdhary, and D. Sengupta. Focs: Fast overlapped community search. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2974–2985, Nov 2015.
3. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003.
4. M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat. Local descriptors in application to the aging problem in face recognition. *Pattern Recognition*, 46:2634–2646, 2013.
5. C. X. C, D. Guanzhong, and Y. Libing. Survey on spectral clustering algorithm. *Computer Science*, 35:14–18, 2008.
6. F. Dornaika, A. Moujahid, Y. E. Merabet, and Y. Ruichek. Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors. *Expert Systems with Applications*, 58:130 – 142, 2016.
7. C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, Aug 2013.
8. S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75 – 174, 2010.

9. Garima, H. Gulati, and P. K. Singh. Clustering techniques in data mining: A comparison. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 410–415, March 2015.
10. Y. C. Gong and C. C. L. spectral clustering. *Advances in Artificial Intelligence*, chapter Locality spectral clustering, pages 348–354. Springer Berlin Heidelberg, 2008.
11. H. Hu. *Graph Based Models for Unsupervised High Dimensional Data Clustering and Network Analysis*. PhD thesis, University of California, 2015.
12. D. Huang, C. Shan, M. Ardabilian, and Y. Wang. Adaptive particle sampling and adaptive appearance for multiple video object tracking. *IEEE Trans. on Systems, Man, and Cybernetics-Part C: Applications and reviews*, 41(6):765–781, November 2011.
13. B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.
14. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.
15. M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
16. R. Nock and F. Nielsen. Statistical region merging. *IEEE Trans. Pattern Anal. Mach*, vol. 26, no 11:1452–1458, 2004.
17. T. Ojala, M. Pietikäinen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.
18. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
19. T. Sajana, C. M. S. Rani, and K. V. Narayana. A survey on clustering techniques for big data mining. *Indian journal of Science and Technology*, 9(3):1–12, 2016.
20. S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):2007, 2007.
21. P. Sharma and J. Suji. A review on image segmentation with its clustering techniques. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(5):209–218, 2016.
22. R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning. Parsing natural scenes and natural language with recursive neural networks. In *International Conference on Machine Learning*, 2011.
23. C. Sousa, S. Rezende, and G. Batista. Influence of graph construction on semi-supervised learning. In *European Conference on Machine Learning*, pages 160–175, 2013.
24. V. Takala, T. Ahonen, and M. Pietikäinen. Block-based methods for image retrieval using local binary patterns. In *Image Analysis, SCIA*, volume LNCS, 3540, 2005.
25. O. Tuzel, F. Porikli, and P. Meer. A fast descriptor for detection and classification. In *European Conf. on Computer Vision*, pages 589–600, 2006.
26. G. Wacquet, E. P. Caillault, D. Hamad, and P. A. Hebert. Constrained spectral embedding for k-way data clustering. *Pattern Recognition Letters*, 34(9):1009–1017, 2013.
27. X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014.
28. L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Faces in Real-Life Images Workshop in ECCV*, 2008.
29. D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.
30. P. Yang, Q. Zhu, and B. Huang. Spectral clustering with density sensitive similarity function. *Knowledge-Based Systems*, 24:621–628, 2011.
31. L. Zhang, S. Chen, and L. Qiao. Graph optimization for dimensionality reduction with sparsity constraints. *Pattern Recognition*, 45:1205–1210, 2012.
32. L. Zhang, L. Qiao, and S. Chen. Graph-optimized locality preserving projections. *Pattern Recognition*, 43:1993–2002, 2010.

Multiview Video Coding: A Comparative Study Between MVC and MV-HEVC

Seif Allah El Mesloul Nasri, Abdul Hamid Sadka, Noureddine Doghmane, and Khaled Khelil

1 Introduction

During the last few years, the demand for higher-resolution video has witnessed a steady increase as much as the demand for interactive and three-dimensional (3D) visual content. It is predicted that the video traffic on the Internet will occupy 82% of all transmitted data by 2021 [1], and the 3D video content with its different formats will indeed be part of this traffic. 3D video content is today not only used for entertainment and leisure but it is also applied in several critical domains such as education, surveillance, cultural heritage and medicine [2]. The multiview video format offers a 3D experience to the end user, through depth sensation in addition to motion parallax. At least two cameras capture the multiview video from slightly different view angles. There exists a considerable amount of inherent redundancy between the viewpoints of the multiview video. Consequently, inter-view coding has been proposed taking into account the resemblance between the recorded views. Recent video coding standards such as H.264 [3] and H.265 [4] provide extended profiles that take advantage of the inter-view resemblances for better compression

S. A. El. M. Nasri
Department of Electronic and Computer Engineering, Brunel University, London, UK
e-mail: seif.nasri@brunel.ac.uk

Department of Electronics, University Badji Mokhtar, Annaba, Algeria

A. H. Sadka (✉)
Department of Electronic and Computer Engineering, Brunel University, London, UK
e-mail: abdul.sadka@brunel.ac.uk

N. Doghmane
Department of Electronics, University Badji Mokhtar, Annaba, Algeria

K. Khelil
LEER Lab, University Mohamed-Cherif Messaadia, Souk Ahras, Algeria

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_7

efficiency. Based on the exploitation of both temporal and inter-view prediction, Merkle et al. proposed an approach that ensures a good trade-off between the bit rate and the video quality [5]. It was adopted and implemented by the Joint Video Team of ISO/IEC, Moving Picture Experts Group and ITU-T, and Video Coding Experts Group in a reference model named the Joint Multiview Video Model (JMVM) [6] or simply MVC, which is the extended profile of AVC/H.264. High compression efficiency is still the main requirement for multiview video coding in addition to other specific requirements such as low-delay random temporal and view access. Many research efforts [7–11] based on the MVC standards have been made with view of improving the coding capability with regard to the multiview video requirements list [12].

The first edition of the High Efficiency Video Coding standard (H.265) was finalised in 2013 by the Joint Collaborative Team on Video Coding (JCT-VC). The H.265 standard can achieve 50% bit rate saving for an equal perceptual video quality compared to H.264 [4]. Back in July 2012, the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) was established by the ISO/IEC MPEG and ITU-T Video Coding Experts Group (VCEG) in order to develop the next generation of the 3D video coding standards. As a result, the second edition with scalability extension (SHVC) [13] and multiview extension (MV-HEVC) [14] was completed in 2014 and published in early 2015.

In this manuscript, an evaluation of the MV-HEVC coding is presented in terms of compression efficiency relative to MVC. The assessment was conducted using multiple multiview video sequences with different contents and qualities.

The remainder of this chapter is organised as follows. Section 2 describes the multiview video and its coding principles. MV-HEVC technical concepts and features are presented in Sect. 3. Section 4 reports the compression performance of the two compared extensions. Finally, the conclusion is given in Sect. 5.

2 MVC Background

Multiview video can be produced when a set of synchronised cameras capture the same scene. The cameras record different angles of the same scene with an enriched overlapped content. Thus, 3D information of the scene is generated based on the cameras' similar content, offering an enhanced visual experience through depth feeling and motion parallax.

Multiview video visualisation is possible through a flat panel screen employing either parallax barrier or lenticular sheet technology (Fig. 1) to perceive a 3D image when both of the viewer's eyes are anywhere within the viewing zone. This so-called autostereoscopic display can support multiple viewers, each seeing 3D from his or her point of view. Looking around objects in the scene simply needs moving the viewer's head. More types of autostereoscopic displays are detailed in [15].

Typically, the conventional two-dimensional (2D) video is formed of continuous groups of pictures with a frequency of (25, 30 or 60 . . . etc.) frames per second.

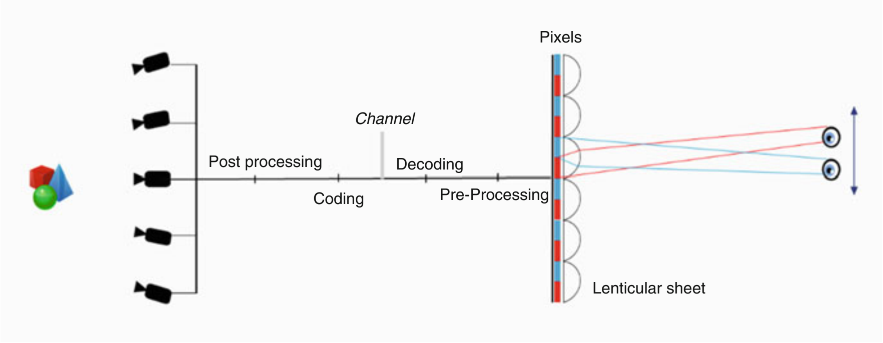


Fig. 1 Multiview video system

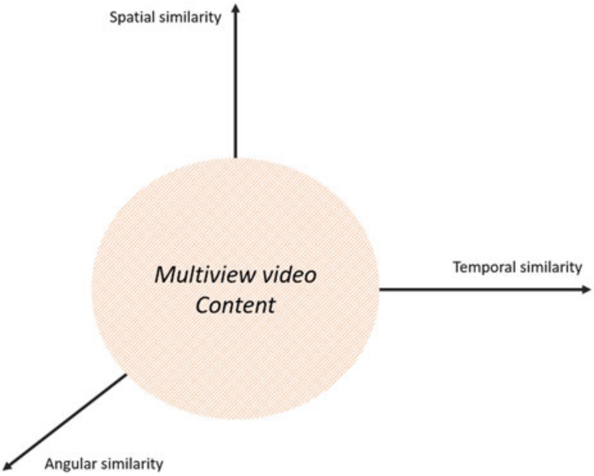


Fig. 2 Multiview video content similarities diagram

Successive frames have a certain degree of similarity, for example, an action video will have less similarity between its successive frames compared to an official speech clip with a fixed background. The same remark is noted for the temporal level between the successive frames [16]. This is associated with the inter-view correlation that exists among the views of the multiview video. The three types of similarity represent, in fact, a redundant information to be exploited to improve the compression efficiency. Figure 2 depicts the 3D similarity that exists within a multiview video content.

The synchronised sequences of the multiview video are coded jointly and simultaneously by only one video codec which is the MVC, as shown in Fig. 3. The MVC has to employ algorithms and techniques to reduce the amount of redundant data during the compression process.

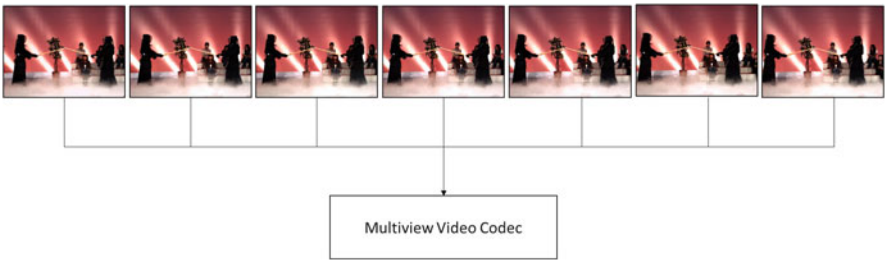


Fig. 3 Multiview video codec input

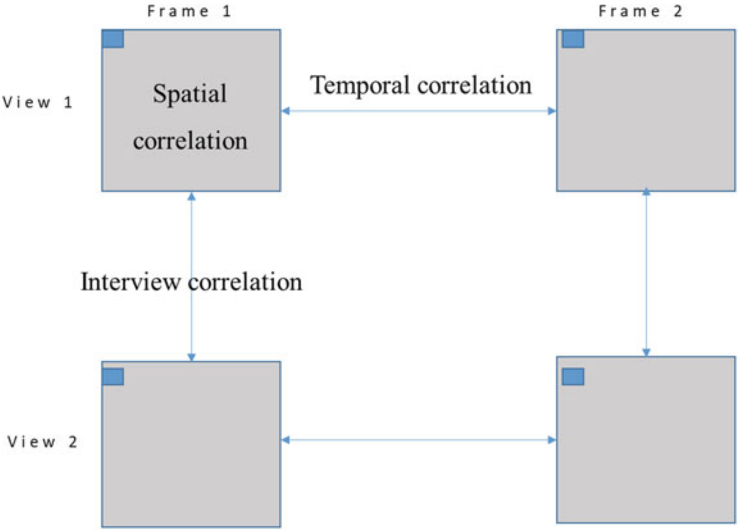


Fig. 4 Redundant information exploitation in the multiview video content

Effective motion-compensated and disparity prediction algorithms are used to eliminate the redundant information between the successive frames and adjacent views, respectively (Fig. 4).

The simplest method for coding a multiview video is the simulcast method, which performs the compression by exploiting only the spatiotemporal redundancies and coding each view independently using a conventional video codec. By making use of H.264/AVC and the hierarchical B pictures, video compression has been efficiently improved in comparison to the traditional simulcast coding structures [17]. Figure 5 depicts the hierarchical B pictures structure where the number of frames in the group of pictures (GOP) is equal to 8. The first picture is independently coded as an instantaneous decoder refresh (IDR) picture, and the

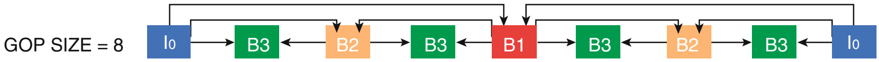


Fig. 5 Hierarchical B pictures structure

so-called anchor or key pictures are coded within regular intervals. The B pictures, located between two I pictures and known as non-key frames, are hierarchically predicted using the concept of hierarchical B pictures.

Despite the fast random access provided by the simulcast coding method, its coding efficiency is not optimal as it neglects the inter-view dependencies during the compression process. Simulcast method is typically employed as a reference model for coding performance comparisons between different MVC schemes.

In fact, research on multiview video coding has been active for more than 30 years since the emergence of the disparity compensated concept in 1986 [18], followed by other propositions in 1989 [19] and 1992 [20]. The first official standardisation of the MVC was in 1996 [21] and consisted of extending H262/MPEG-2 [22] capabilities to support the multiview video content. However, at that time, the ultimate challenge was to upgrade video services from the standard analogue definition to the digital high definition. This fact prevented the multiview extension of H.262/MPEG-2 from being applied and developed.

Following the progress in video compression technologies and multimedia services, MPEG launched a call for proposal on MVC in July 2005. Based on the AVC/H.264 coding standards, some proposed responses introduced different forms of inter-view prediction structures [23–25]. Compared to the simulcast coding where each view is coded independently, the inter-view coding methods offer significant gains in terms of bit rate saving. Merkle et al.'s [5] approach was adopted and implemented by the Joint Video Team in a reference model named the Joint Multiview Video Model (JMVM) [6].

Figure 6 presents the inter-view prediction structure used as the default structure of JMVM. Eight views (cameras) are employed in this scheme where S_n indicates the different cameras, while T_n represents the time location of the frames. Moreover, in this case, each group of groups of pictures (GGOP) is composed of eight views and eight pictures per GOP.

The IBP structure employs three types of views: I-view as one base view per Group of groups of pictures, P-views which are predicted from a unique direction and B-views involving bi-directional inter-view prediction for coding its set of frames. Much research work based on MVC/H.264 has been undertaken to improve the outcomes of the coding process in terms of view random access [26, 27]. Meanwhile, research on MVC/AVC is still underway while HEVC codec implementation in the market is going at slow pace due to its loyalty cost and complexity.

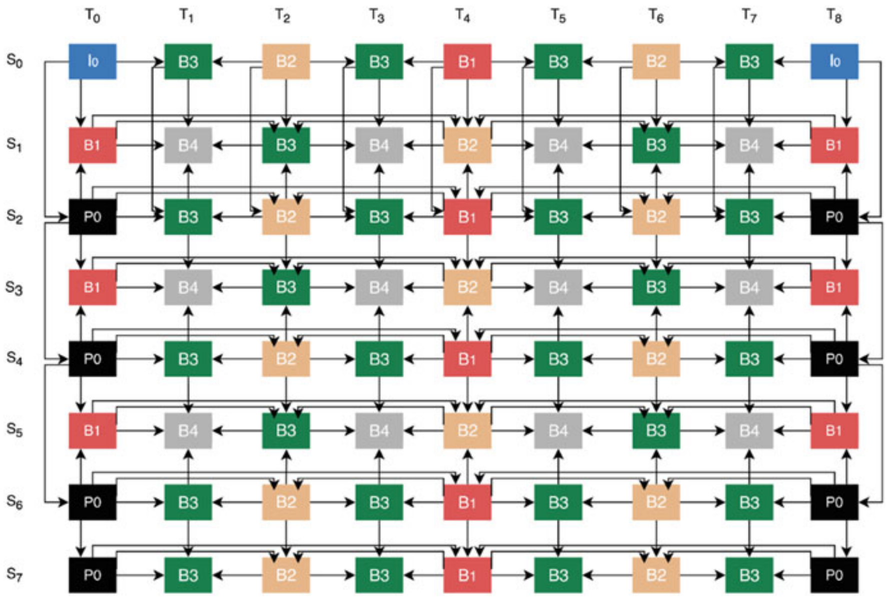

T_0 T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8
S_0 I0 B3 B2 B3 B1 B3 B2 B3 I0
S_1 B1 B4 B3 B4 B2 B4 B3 B4 B1
S_2 P0 B3 B2 B3 B1 B3 B2 B3 P0
S_3 B1 B4 B3 B4 B2 B4 B3 B4 B1
S_4 P0 B3 B2 B3 B1 B3 B2 B3 P0
S_5 B1 B4 B3 B4 B2 B4 B3 B4 B1
S_6 P0 B3 B2 B3 B1 B3 B2 B3 P0
S_7 P0 B3 B2 B3 B1 B3 B2 B3 P0

Fig. 6 IBP prediction structure

3 MV-HEVC

Results of subjective evaluation [28] show that HEVC/H.265 standard can reach the same quality levels as H.264/AVC whilst generating approximately 50% lower bit rate on average. HEVC standard adopts innovative tools which contribute to achieving this gain, such as accurate intra-/inter-predictions, in-loop sample adaptive offset filter and quadtree-based block partitioning [4].

HEVC benefits from using variable pattern comparison and difference-coding areas starting from blocks of 16 × 16 to 64 × 64 pixels. The concept behind this is based on partitioning the frame into coding tree units (CTUs), which replace the macroblocks used in H.264. Each CTU contains two chroma and one luma coding tree blocks CTBs. CTB size can be 16 × 16, 32 × 32 or 64 × 64, where larger pixel block size increases the compression efficiency. The CTBs are then divided into one or more coding units (CUs) as shown in Fig. 7. The CU is split into prediction units (PUs), a basic entity for intra- and inter-predictions, variable in size from 64 × 64 to 4 × 4 pixels. Variable partition scenarios have been defined in the design of the HEVC encoder considering a certain attention to complexity. For instance, to deal with critical case memory bandwidth in the decoding process, PUs coded using temporal inter-prediction are restricted to the minimum size of 8 × 8 if they are bi-predicted from two references, or 8 × 4 or 4 × 8 if they are predicted from a single reference [4].

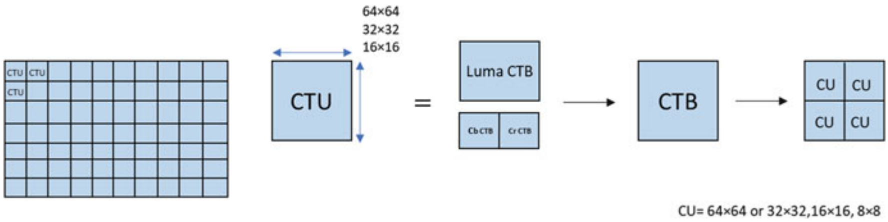


Fig. 7 Block partitioning in HEVC

Compared to AVC which includes only eight directional modes for the intra-picture prediction, HEVC employs 33 intra-picture prediction modes in addition to planar (Mode 0, surface fitting) and DC (Mode 1, flat) prediction modes. Due to the increased number of modes (35), efficient coding of intra-prediction mode is achieved by using a list-based approach. For each prediction unit, the most probable three modes are determined and a Most Probable Mode (MPM) list is constructed from these modes.

The HEVC bitstreams include an elementary unit called a network abstraction layer (NAL) unit, composed of payload and a header. The NAL header consists of a 5-bit NAL unit type, 6-bit layer identifier called nuh_layer_id, and a 3- bit temporal sub-layer identifier. A new video parameter set (VPS) structure has been included in HEVC as metadata representation to allow the extension compatibility of the standard including dependences between temporal sub-layers. It also contains essential data that can be shared with the decoding process.

The multiview extension of HEVC uses the same fundamental coding tools of the HEVC main profile in addition to some specific features mainly related to the stereoscopic and multiview representations. MV-HEVC provides bit rate saving compared to the standard HEVC simulcast by enabling the exploitation of the inter-view references within the motion-compensated prediction. It is also noted that MV-HEVC utilises the same coding design principle (IBP) as the multiview extension of H.264.

However, the concept of inter-view has been replaced in MV-HEVC by the inter-layer prediction design. The multi-layer approach is employed in all multi-layer extensions [29], including MV-HEVC, 3D-HEVC, as well as the scalable extension of HEVC (SHVC). A layer can represent a depth, texture or other auxiliary information related to a particular camera view. All layers of the same camera perspective are marked as a view; while layers representing the same type of information are denoted as components in 3D video (Fig. 8).

MV-HEVC includes high-level syntax (HLS) additions [14] and can be implemented using existing 2D single-layer decoding cores. Moreover, MV-HEVC shares the same HLS with all HEVC multilayer extensions. HLS enables the extraction of a single texture base view from MV-HEVC bitstream which is decodable by the main profile HEVC decoder.

Figure 9 shows an example of MV-HEVC bitstream with three texture views coded by the so-called IBP inter-view structure. The base layer (left view) is coded

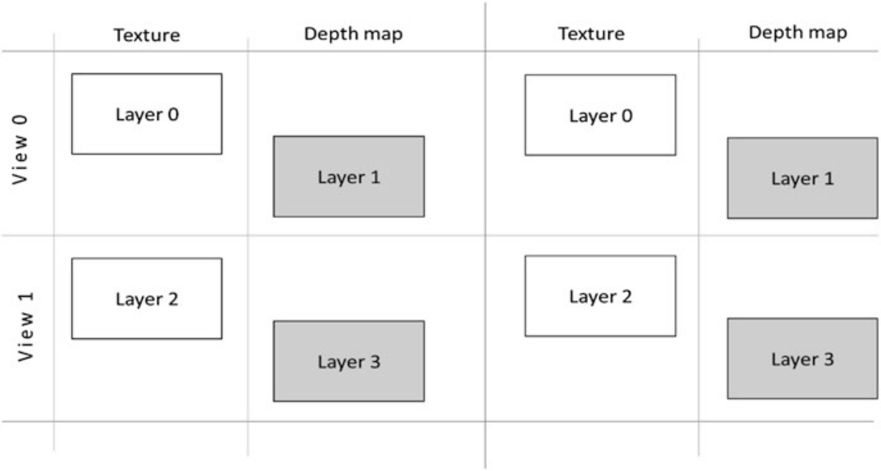


Fig. 8 Layers division in MV-HEVC

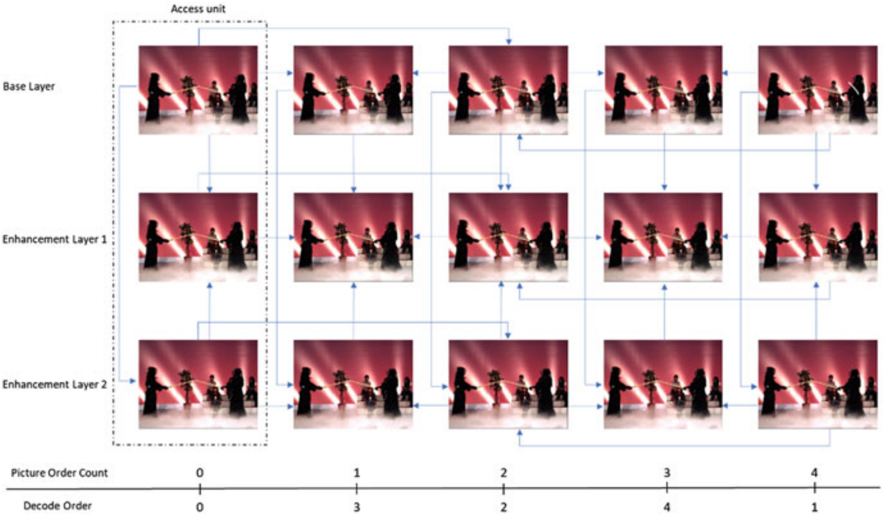


Fig. 9 MV-HEVC bitstream with three texture views using IBP inter-view prediction

independently of other views using HEVC main profile. The MV-HEVC profile is enabled to code the two enhancement layers (ELs). EL2 (right view) utilises inter-view prediction from the base layer, and EL1 (centre view) is predicted from both left and right views.

4 Experimental Results

In this section, the compression efficiency of MV-HEVC and MVC is compared and evaluated. Four different video sequences have been used in the experiments. Table 1 describes the used multiview video sequences and their parameters. Also, samples of the tested sequences are illustrated in Fig. 10.

Table 1 Multiview video sequences used for the compression efficiency evaluation

Database	Video sequences	Frame rate	Image resolution	Camera parameters
MERL	Vassar	25	640 × 480	8 cameras/20 cm spacing
MERL	Ballroom	25	640 × 480	8 cameras/20 cm spacing
Fujii Lab	Kendo	30	1024 × 768	7 cameras/5 cm spacing
Fujii Lab	Balloon	30	1024 × 768	7 cameras/5 cm spacing

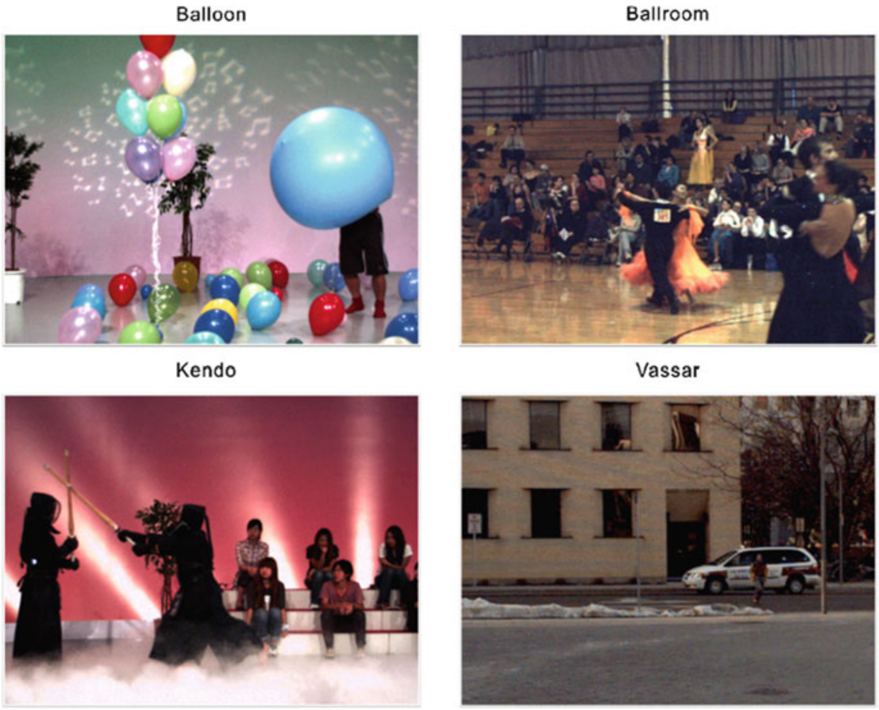


Fig. 10 First view picture of the used multiview video sequences

Table 2 Initial common encoding configuration

Frames to be encoded	250
GOP size	8
Intra period	8
Quantization Parameter	[25,30,35,40]
Search mode	Fast mode
Search range	64

The objective evaluation is shown using graphs of peak signal-to-noise ratio PSNR (dB) versus bit rate (kbit/s). The PSNR which expresses the video quality is given by:

$$\mathrm{PSNR} = 10 \times \log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right) \quad (1)$$

MSE represents the mean square error between the original and the compressed video signals. Conventionally, the objective measure of quality is applied to the luminance video signal regardless of the chrominance signals. Table 2 regroups the common primary conditions that have been used to obtain a fair comparison. The quantisation parameter (QP) controls the quality of the compressed video and the bit rate of the generated bitstream; the higher the value of the QP, the lower is the bit rate and the video quality. Four QP values are chosen according to the standardisation tests defined in [30].

It can be clearly inferred from Figs. 11 and 12 that the MV-HEVC exceeds the MVC in terms of bit rate saving and video quality. This outperformance ultimately covers all the conducted tests through the different datasets and conditions. The rate distortion (RD) curves of the high-definition multiview video sequences, shown in Fig. 11, prove that MV-HEVC codec improves the compression performance compared to MVC over the entire bit rate range. For instance, for QP = 25, the bit rate saving gain achieved by MV-HEVC exceeds 25% and 31% for Balloon and Kendo sequences, respectively. Furthermore, Fig. 12 shows that the MV-HEVC bit

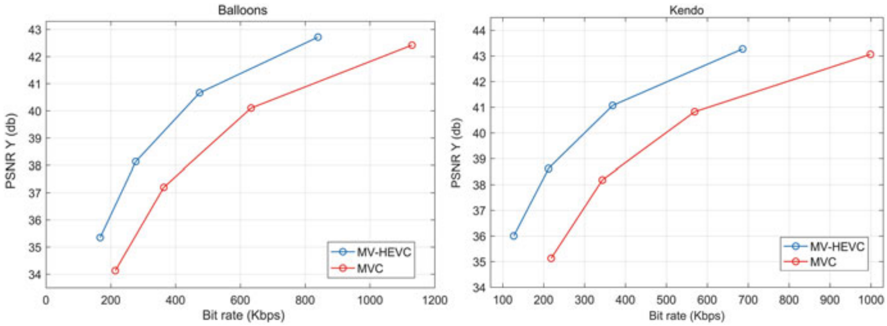


Fig. 11 Compression efficiency comparison through HD multiview video sequences

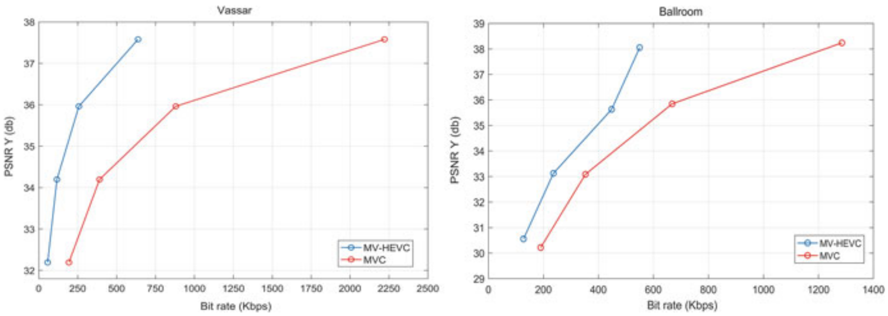
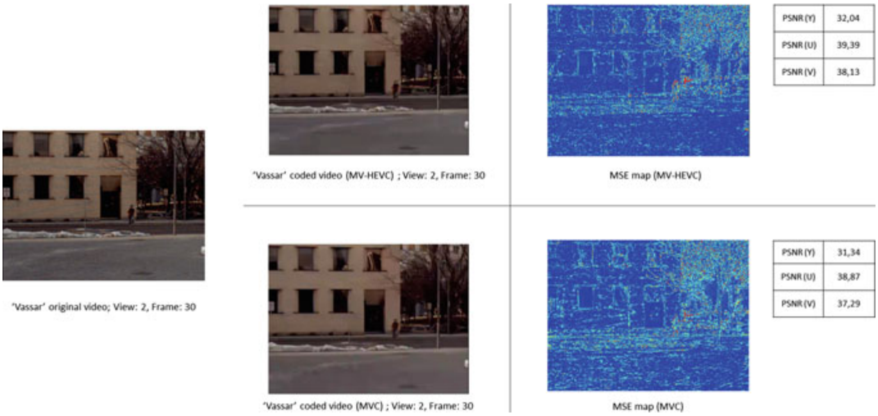


Fig. 12 Compression efficiency comparison through SD multiview video sequences


'Vassar' coded video (MV-HEVC) ; View: 2, Frame: 30
MSE map (MV-HEVC)
'Vassar' original video; View: 2, Frame: 30
'Vassar' coded video (MVC) ; View: 2, Frame: 30
MSE map (MVC)

PSNR (Y)	32,04
PSNR (U)	39,39
PSNR (V)	38,13

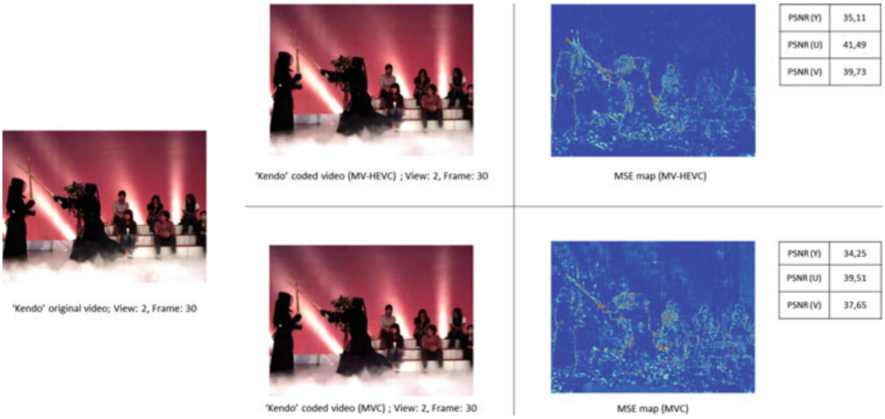
PSNR (Y)	31,34
PSNR (U)	38,87
PSNR (V)	37,29

Fig. 13 Image quality comparison between MV-HEVC and MVC using Vassar sequence

rate saving is further increased for the standard definition sequences, whereby a gain of 71% and 57% is achieved for Vassar and Ballroom sequences, respectively.

Figures 13 and 14 present a frame-based comparison between MV-HEVC and MVC codecs. Frame number 30 located in view 2 (camera 2) of the two chosen multiview video sequences is selected for this comparison. This frame, which comes after three successive groups of pictures, is coded using both temporal and inter-view predictions. Also, the quantization parameter QP = 40 has been selected for this comparison to evaluate the performance of the reported codecs at the lowest level of perceptual image quality.

Figure 13 shows the comparison using a standard resolution video (Vassar), the degradation can be seen in the compressed frame with MV-HEVC and MVC as well. However, the difference cannot be clearly perceived between the two compressed frames. The MSE maps slightly highlight the difference between the two compressed frames, where extra red regions are observed in the frame compressed by MVC codec, which indicates a larger number of mismatching errors.


'Kendo' original video; View: 2, Frame: 30
'Kendo' coded video (MV-HEVC) ; View: 2, Frame: 30
MSE map (MV-HEVC)
'Kendo' coded video (MVC) ; View: 2, Frame: 30
MSE map (MVC)

PSNR (Y)	35,11
PSNR (U)	41,49
PSNR (V)	39,73

PSNR (Y)	34,25
PSNR (U)	39,51
PSNR (V)	37,65

Fig. 14 Image quality comparison between MV-HEVC and MVC using Kendo sequence

However, the blue regions, which represent the matching between the original and the compressed frames, are distinctly perceived in the frame compressed by MV-HEVC. The PSNR values confirm the MSE map results, where the PSNR (Y) gain of MV-HEVC is 0.7 dB, and the overall value is 0.69 dB.

Almost a similar perception can be obtained from Fig. 14 where HD multiview video sequences have been used with the same quantisation parameter value for both codecs. The results emphasize the same fact that MV-HEVC outperforms MVC in terms of image quality with a gain of 0.86 dB achieved for PNSR(Y) and 1.64 dB for the mean value which includes PSNR(Y), PSNR(U) and PSNR(V).

5 Conclusion

The chapter reviewed the multiview video coding theory and concepts, focusing on MVC and MV-HEVC coding standards. Both codecs use the same IBP design for the disparity compensation in addition to the hierarchical B algorithm for the temporal level. The MV-HEVC employs the powerful tools of HEVC such as the innovative block partitioning to improve the rate distortion capability. Both codecs have been implemented and evaluated through different datasets and common test conditions. The used test video sequences were multiple texture views without depth map of SD and HD resolutions. Test results have shown an increased compression efficiency of MV-HEVC compared to MVC. The significant bit rate saving gain starts from 24% for Balloon sequences and achieves 70% for Vassar sequences.

Acknowledgement This publication was made possible by NPRP grant 9-181-1-036 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Cisco, Visual Networking Index: Forecast and Methodology, 2016–2021 June 6, 2017.
2. A. Smolic et al., "3D video and free viewpoint video—technologies, applications and MPEG standards", IEEE Int. Conf. Multimedia and Expo, pp. 2161–2164, IEEE, Toronto, Ontario, Canada (2006).
3. J. Ostermann et al., "Video coding with H.264/AVC: tools, performance, and complexity", IEEE Circuits and Systems Magazine. 4(1), 7–28 (2004)
4. G. J. Sullivan et al., "Overview of the High Efficiency Video Coding (HEVC) Standard", IEEE Circuits and Systems Magazine. 22(12), 1649–1668 (2012).
5. P. Merkle et al., "Efficient prediction structures for multiview video coding," IEEE Trans. Circuits Syst. Video Technol. 17(11), 1461–1473 (2007).
6. Y. Chen, P. Pandit, and S. Yea, "WD 4 reference software for MVC," ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/SG16, Doc. JVT-AD207 (2009).
7. Y. Yang, Q. Dai, J. Jiang, and Y-S. Ho, "Coding order decision of B frames for rate-distortion performance improvement in single-view video and multiview video coding," IEEE Trans. Image Processing, vol. 19, no. 8, Aug. 2010.
8. S.H. Hany, M. El-Khamy and M. El-Sharkawy, "Blind configuration of multi-view video coder prediction structure", IEEE Trans. Consumer Electron, vol. 59, no. 1, pp. 191–199, 2013
9. A. B. Ibrahim, A. H. Sadka "Error resilience and concealment for Multiview video coding" IEEE Int. Symp. on Circuits and Systems 2014, pp. 1–5 (2014).
10. Pei-Jun Lee, Ho-Ju Lin, and Kuei-Ting Kuo, "Faster mode determination algorithm using mode correlation for multi-view video coding," IET Signal Process, vol. 8, no. 5, pp. 565–578, 2014.
11. S. Nasri et al., "Enhanced view random access ability for multiview video coding," J. Electron. Imaging 25(2), 023027 (2016).
12. "Requirements on multi-view video coding v.4," ISO/IEC JTC1/SC29/WG11, Doc. N7282, Poznan, Poland (2005).
13. J. M. Boyce et al., "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding Standard", IEEE Trans. Circuits Syst. Video Technol. 26(1), 20–34 (2016).
14. G. Tech et al., "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding", IEEE Trans. Circuits Syst. Video Technol. 26(1), 35–49 (2016).
15. N.A. Dodgson, "Autostereoscopic 3D displays", Computer, 38(8), 31–36 (2005).
16. A. H. Sadka, Compressed Video Communications. Halsted Press, 2002.
17. L. Ma and F. Pan, "Efficient compression of multi-view video using hierarchical B pictures," in Int. Conf. on Multimedia and Ubiquitous Engineering, pp. 118–121 (2008)
18. M. E. Lukacs, "Predictive coding of multi-viewpoint image sets", IEEE Int. Conf. Acoust. Speech Signal Process, vol. 1, 521–524. Tokyo, Japan (1986).
19. I. Dinstein et al., "On the compression of stereo images: Preliminary results", Signal Process., Image Commun., 17(4), 373–382 (1989).
20. M. G. Perkins "Data compression of stereo pairs", IEEE Trans. Commun., 40(4), 684–696, (1992).
21. ITU-T and ISO/IEC JTC 1, Final draft amendment 3, Amendment 3 to ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video), ISO/IEC JTC 1/SC 29/WG 11 (MPEG) Doc. N1366 (1996).
22. ITU-T and ISO/IEC JTC 1, Generic coding of moving pictures and associated audio information Part 2: Video, ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video) (1994).
23. MPEG Video Sub-Group Chair (J.-R. Ohm), Submissions received in CfP on multiview video coding, Bangkok, Thailand, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) Doc. M12969 (2006).
24. MPEG Video and Test Sub-Groups, Subjective test results for the CfP on multi-view video coding, Bangkok, Thailand, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) Doc. N7799 (2006).
25. K. Muller et al., "Multiview coding using AVC", Bangkok, Thailand, ISO/IEC JTC 1/SC 29/WG 11 (MPEG) Doc. M12945 (2006).

26. A. Bekhouch et al., “Improving view random access via increasing hierarchical levels for multi-view video coding”, IEEE Transactions on Consumer Electronics, 62(4), 437–445 (2016).
27. S. Nasri et al., “Group of Pictures Effects on Proposed Multiview Video Coding Scheme”, 40th International Conference on Telecommunications and Signal Processing, pp. 548–554 (2017).
28. J.-R. Ohm et al., “Comparison of the coding efficiency of video coding standards— Including High Efficiency Video Coding (HEVC)”, IEEE Trans. Circuits Syst. Video Technol., 22(12), 1669–1684 (2012).
29. R. Sjoberg et al., “Overview of HEVC high-level syntax and reference picture management”. IEEE Trans. Circuits Syst. Video Technol., 22(12), 1858–1870 (2012).
30. V. Baroncini et al., “MV-HEVC Verification Test Report”, Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-N1001 (2016)

Part II
Data Handling and Management (Including Data Security, Database Handling, Cloud Computing, Hardware and Software Technologies)

Data Fragmentation Scheme: Improving Database Security in Cloud Computing

Amjad Alsirhani, Peter Bodorik, and Srinivas Sampalli

1 Introduction

NIST [1] defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." Cloud computing involves storing data using a third-party or noncentral storage mechanism and requires the ability to access this data from anywhere at any time. It offers many services and includes a variety of models to meet users' needs at affordable prices. One of the cloud computing characteristics is scalability, whereby data is scaled around the cloud providers' servers. The robust devices cloud providers rely on to operate the cloud give users fast network speeds, high performance processing, and a vast amount of storage space.

Despite the perceived benefits of cloud computing, there are still significant security concerns surrounding storing data in the cloud. A standout among these concerns is adequately protecting the confidentiality of sensitive data. This ongoing and highly important concern has motivated us to investigate a means to enhance security in a cloud computing context and to create a viable approach to provide security to data stored in cloud. Cloud computing has many attractive advantages that encourage potential users to consider moving to a modern style of computing.

This is an extended version of a preliminary conference paper that was accepted and presented in ICCA 2017 [2].

A. Alsirhani · P. Bodorik (✉) · S. Sampalli
Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
e-mail: amjada@cs.dal.ca; bodorik@cs.dal.ca; srini@cs.dal.ca;
http://web.cs.dal.ca/~bodorik/

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_8

However, these benefits, as Hacg et al. [3] claim, come at a high price, with users facing more significant privacy risks and increased vulnerabilities when they move their, often private or sensitive, data to a cloud. More importantly, concerns arise about data confidentiality, because the data is often not under the direct control of the owner. Consequently, providers can compromise and access sensitive data, which constitutes an invasion of the database owner's privacy. As many cloud computing providers may not even trust their employees, cloud clients find it a challenging assignment to find trusted providers to store sensitive data [4]. Additionally, in some cases, the terms and conditions of cloud computing services are subject to any changes by the providers as they reserve the right to do so. Therefore, the data privacy and confidentiality risk is essential to consider [5].

Considering ongoing security-related issues, encryption rises as the most straightforward solution wherein the data is being encrypted before sending it to the cloud and thus preventing providers from obtaining sensitive information. Unfortunately, however, encrypted data cannot be regularly queried, making it hard for users to retrieve searchable data. Some proposed solutions to this dilemma suggest using asymmetric (i.e., public key) cryptography, where the key is being shared with the cloud providers. Nevertheless, the provider can still infer sensitive information to perform the decryption in response to the client's query.

Similarly, symmetric key (i.e., secret key) cryptography involves decryption on the provider's side, allowing providers to derive sensitive information in this scenario too. Consequently, neither secret key nor public key cryptography offers a suitable solution. Because there is no individual encryption algorithm that can support all Structured Query Language (SQL) queries without decryption, there is a demand for a system that provides users with security while also supporting a variety of query types. This leads us to the following questions: How can we guarantee data confidentiality while using untrusted cloud computing provider resources, and what encryption algorithms can be utilized to support a variety of queries?

2 Related Work

Work regarding the confidentiality of data stored through outsourcing is categorized into four groups: **(a)** fragmentation scheme (a column-based partition), **(b)** hardware-software solution, **(c)** sensitive data vs. nonsensitive data, and **(d)** combination of encryption algorithms. More details about these methods are provided below.

(a) Hacig and Li [6] proposed a fragmentation scheme, which is a column-based partition. On the server side, there are only vertical fragments that are encrypted. A query involving the fragment's data results in fetching the fragment from the cloud to the client and decrypting it, after which the query is executed on the fetched and decrypted fragment. Consequently, this approach is not an applicable solution for this problem because it eliminates cloud computing's main benefit in terms of providing storage. It also requires much overhead time to execute the query. Other studies have considered this approach such as [3, 7, 8]. They essentially focus on the query optimization technique as a means of attempting to tackle the limitation of [6], which concerns performance.

(b) Bouganim and Pucheral[9] have proposed a combination of hardware and software solution for securing the data in the cloud computing environment. They claim that a standalone software solution is not guaranteed, because of Internet security vulnerability. Their idea is based on a smart card that works as a mediator between the cloud and the user with an assumption of secure communication between the user and the cloud. The smart card is encrypting/decrypting the data at inserting/retrieving time. The cloud has no control over the smart card, so it is only used to store encrypted data. This approach limits the benefit of the cloud computing because the smart cart does the process of encryption and decryption. Therefore, the limited computational power of smart card led to inefficient encryption and decryption performance. Consequently, this solution is not practical for protecting data stored in a cloud.

(c) Anciaux et al. [4] have proposed visible and hidden data concept to ensure data confidentiality. The data is split into two parts (i.e., sensitive data and nonsensitive data). The sensitive data has to be encrypted and stored in private place without public accessibility. A smart USB key mechanism is used to protect private information. In contrast, the nonsensitive data can be stored in the cloud in plaintext. The two parts of the data are joined for querying when the holders of the USB key plug it into their machine while using a distributed method for joining data. This approach has a scalability limitation because of sensitive data storage technique. Moreover, it also limits the benefits of cloud computing by storing sensitive data at the client side (using a smart USB key). Hence, this solution is not practical especially if the system deals with big data. Other studies belonging to this category can be found in [10–14].

(d) Popa et al. [15] have proposed a practical solution to improve the outsourced data confidentiality level from curious cloud providers. Their scheme involves a number of components, including encryption algorithms, proxy, and user's application. The concept behind the combination relies on the case that no existing encryption algorithm can assist all types of queries, so the authors examined encryption algorithms that support queries to be issued to encrypted data. They considered six encryption algorithms that can be used to support the fundamental query structure. Other studies belonging to this category can be found in [16–20].

3 Methodology

We propose a combination of encryption algorithms and a fragmentation technique that together form a novel contribution to this research [2]. The utilized encryption algorithms are described in Sect. 3.1. Figure 1 presents the architecture of the proposed scheme which is a hybrid cloud that mainly consists of two parts: public

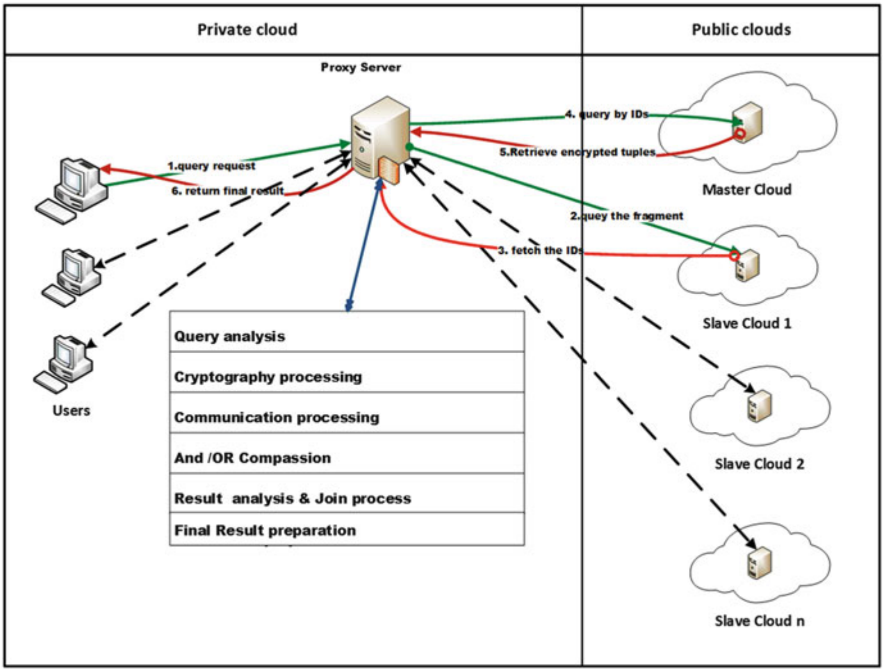


Fig. 1 The architecture of our scheme

clouds and a private cloud. The public clouds include a master cloud, which includes an encrypted copy of the whole database, and *N* slave clouds, which store extended columns/fragments. It is assumed that both the master and slaves are operated by different cloud providers. The public clouds should not know or be able to infer any information about any other fragments of the database stored on other providers' clouds. The concept is to improve security by concealing this information among the clouds. Of course, this assumption causes problems with scalability regarding the number of columns as ideally each of the columns should be stored on a different cloud offered by a different provider and that the providers do not collaborate. In case of a limited number of clouds, the columns can be stored in the same cloud as long as different columns are encrypted using different encryption algorithms with different keys.

The private cloud consists of a proxy and a client system generating queries. The proxy server is located in the private cloud, center of communication, between the client and the public clouds. There is no direct communication between users and public clouds. All communication has to go through the proxy, which translates any received query from the users to a query or a set of queries that are issued to the public clouds. More details are provided in subsequent sections.

3.1 Encryption Algorithms

The utilized encryption algorithms are the same as those used in the approach described in [15]. They include the Advanced Encryption Standard (AES), order-preserving encryption (OPE), homomorphic encryption (HOM), search, and deterministic encryption. In our scheme, the usage is different due to differences in the architecture between the two approaches. The usage and selection of these algorithms are based upon their ability to support users' queries.

3.1.1 Advanced Encryption Standard (AES) Algorithm

The AES algorithm is used to encrypt the entire database in the master cloud, except for the index column. The index column, a system-generated primary key, is used to identify the tuples of the relation and is replicated in each extended column fragment. A slave cloud contains an encrypted column of the base table together with an unencrypted copy of the index column. The keys for encryption/decryption are stored at the proxy server so that encryption and decryption are performed only within the private cloud. Due to the numerous security advantages of the cipher block chaining (CBC) mode, such as confidentiality, it is our best choice for encrypting the master cloud replica.

3.1.2 Order-Preserving Encryption (OPE) Algorithm

The OPE algorithm is one of the fundamental encryption algorithms used in encrypted relational databases. Most databases need comparison operation between values, particularly numerical ones. With this in mind, the OPE algorithm can be applied to encrypted data to provide a comparison operation Liu et al. [21]. The OPE algorithm is based on linear expression as well as random noise. The encryption algorithm is modeled by Eq. (1), where the coefficients a and b are the keys and v is the value that needs to be encrypted.

$$E(v) = a * v + b + \text{noise} \tag{1}$$

The noise part of the expression, which is added to improve security, is a random number in the range of zero and a coefficient a multiplied by sensitivity. The sensitivity value is 0.01 or 1, depending on whether the value is a double or an integer, respectively. The coefficient "a" can be used to control the range of the noise value. Therefore, the algorithm will produce a different ciphertexts for the same plaintext, depending on the noise, which will enhance security because the pattern will be hidden [21].

3.1.3 Homomorphic Encryption (HOM) Algorithm

The HOM encryption algorithm is needed in most encrypted relation databases due to the mathematical properties in that it can support arithmetic operations, such as multiplication, summation, averages, etc. Queries that include mathematical operations can be applied on encrypted data if the HOM algorithm is used for encryption, and hence it is becoming the solution for applications that require mathematical functionality applied on encrypted data. For instance, the result of multiplication of two encrypted data will decrypt to the sum of their corresponding plaintexts. We calculate the sum of values by using Eq. (2):

$$D(E(a) * E(b)) = a + b \tag{2}$$

3.1.4 Search Encryption Algorithm (RC4)

Word search queries are essential for a string column in a database. Many Database Management System (DBMS) support that kind of queries in the context of plaintext database. However, in context of an encrypted database, there also has to be a way to query an encrypted text. Therefore, a deterministic encryption algorithm is used to encrypt string column so that it can be queried [15]. The algorithm (search) was originally proposed in [22] to search for a string token in document files. It is maintaining the letter position as one of the algorithm's inputs. In [15], the algorithms were adapted to support a word search in the context of databases such as searching for someone's name. Serving the same purpose, RC4 is not only a deterministic encryption algorithm, but it is also fast at encrypting and decrypting [23]. Its properties make it one of the best choices for supporting string searches.

3.1.5 Deterministic Encryption Algorithm (DET)

There are many operating modes in which security and performance differ in the AES algorithm [24]. The ECB mode of the algorithm has the idempotence property, making it possible to perform the equality operation upon encrypted data. This algorithm is used to support equality and join operations [15].

Although this algorithm has the weakness of showing patterns; the confidentiality will improve when the fragmentation technique proposed herein is used. Samarati et al. provide an in-depth analysis about how obtaining data from a single column does not reveal any sensitive information [25].

3.2 Fragmentation Technique

In the following text, we discuss a fragmentation technique for a single database relation; it is intended that it be applied for each database relation that requires confidentiality. The fragmentation and replication technique involves two aspects: a master cloud and slave clouds (i.e., column-based vertical fragmentation). In the initial configuration, the entire database is encrypted using a highly secure encryption algorithm and stored in the master cloud while not disclosing the encryption key to the master cloud provider. In fact, none of the keys are disclosed to any of the cloud providers. Furthermore, when a relation for the master cloud is created, we ensure that the relation has one column that is used as an index to the tuples of the relation. Thus, in addition to the primary key of the original relation, we create another attribute (table column) containing a dense index, which contains unique values that uniquely identify the tuples of the relation.

The master cloud is the essential part of our scheme as it maintains the entire relation in one place. Hence, the encryption algorithm to be used in the master cloud needs to be secure enough to hold sensitive data. Many studies have shown that AES-CBC is a secure and reliable algorithm for the outsourced storage of sensitive data [26, 27], and, consequently, several studies have used AES-CBC to store sensitive data [28–30]. Therefore, in our scheme, we also use the AES-CBC encryption algorithm to encrypt the master cloud relation. The index column is the only column stored in plaintext so that search result on slave clouds can be used to retrieve resulting tuples from the slave cloud. The index column serves as a candidate key and is replicated in each fragment that contains an extended column of the relation. The purpose of storing an entire relation in the master cloud (and not in the client cloud) is to obtain the highest advantage of cloud computing by not storing data on the client side. As the master cloud stores encrypted data using the AES-CBC algorithm, it cannot be used by itself for search queries, and for most queries both the master and slave clouds will need to be used. However, there are some simple queries that can be answered by querying only the master cloud, that is, queries that retrieve all or some of the columns of all tuples or search queries based only on the unencrypted index column. Table 1 shows some examples of the two kinds of queries that can be submitted directly to the master cloud database without querying the slave clouds.

The configuration continues with the vertical fragmentation to create a number of encrypted replicas of the columns that are then stored in the slave clouds. Furthermore, before a column is stored in a slave cloud, it is encrypted while not disclosing the encryption key to the cloud provider. Whether a table column is replicated and which encryption algorithm is used is determined by the observed

Table 1 Examples of master cloud queries

#	Query	Result
1	Select * From Table	Return all tuples
2	Select *column_name(s)* From Table	Return entire column(s)
3	Select * From Table *Limit value*	Return number of tuple(s) based on the limit value
4	Select * From Table Where index in *(value_1, value_2 , value_n)*	Return tuple(s) based on the index value

or anticipated access pattern to the relation. We shall simply state that if the access pattern to the relation R is such that column A is frequently used by the queries with a predicate $(R.A\Theta k)$, where Θ is an arithmetic relational operator and k is a constant, then the column is replicated and encrypted using an algorithm applicable for querying using Θ. Furthermore, each encrypted column is also augmented with an index column to create an extended column. Each replicated and encrypted column is stored in a slave cloud.

If distinct queries access a column of the relation with different predicates, a column of the relation may appear more than once in a fragment, but encrypted using different encryption algorithms. For instance, if a column is accessed by two different queries with two respective predicates $(R.A = k1)$ and $(R.A > k2)$, then the column $R.A$ would appear in a fragment twice, once encrypted using the DET algorithm for queries using the predicate $(R.A = k1)$ and once encrypted using the OPE algorithm for queries using the predicates $(R.A > k2)$.

3.3 Proxy's Functionality

The proxy is an essential element of our system, as it does most of the processing. It performs the creation, insertion, encryption, decryption, query parsing, and the retrieval and composition of results. Critically, the proxy server has to be within the private cloud and must communicate with the outside world through a highly secure channel, such as Secure Sockets Layer (SSL).

When a proxy receives a client query, it parses it and transforms it into a set of sub-queries on extended columns stored in the slave clouds. As a proof of concept, we implement the process of querying the slave clouds only in serial fashion, which increases the overall delay in comparison if the slaves were queried in a parallel fashion. This effect increases with the increases in the number of predicates as for each predicate a slave is queried, such that the slaves are queried serially, one following another as opposed to in parallel. Utilizing the serial fashion has no performance effect on a query that retrieves the whole relations, that is, a query that has no predicates—only the master cloud storing the whole encrypted relation is accessed.

Each slave returns to the proxy a set of indices forming the answer of the sub-query it has received. When the proxy receives the results of sub-queries, it performs either the union or intersection algorithm on the indexes if there is more than one where condition/predicate in the query's Where clause. If these conditions are separated by OR, the union algorithm is used; otherwise, the intersection algorithm is used. Once the indices of the final result are formed, the proxy issues a query to the master relation to fetch the tuples that match them. The proxy performs all the encryption and decryption. Before inserting any values that come from the user, the proxy encrypts them before storing them in the clouds. Any values appearing in the query also need to be encrypted by the proxy before querying any slave clouds.

4 Implementation and Evaluation

4.1 Cloud Computing Tools and Configuration

As a first step in proving our concept, we created a public cloud computing account at Rackspace, which offers open-source cloud computing [31]. We then created a number of servers equal to the number of fragments that we need, plus one more for the master cloud. Both servers have almost the same configuration. However, the proxy server has different features. Table 2 shows the features of the servers.

4.2 Evaluation Method

We evaluate our method by determining the total delay using an analytical model and through modeling. In the evaluation, we concentrate on the total delay (in ms) per user SQL request as the user will receive the whole query result in one response. Our method is compared to two base methods:

(a) The first is an Unsecure method in which the DB is stored in one cloud, and there are no efforts to provide confidentiality. Thus, data is stored and communicated in plaintext. We refer to this method as the Unsecure approach.

(b) The second method is based on an approach appearing in [15], in which the data is encrypted and stored in one cloud. In that cloud, each column is encrypted using the encryption algorithms that support the desired operation(s). We refer to this approach as the Secure Centralized approach. Our method will be referred to as the Secure Distributed Approach (SDA). Experiments report delays due to communication, crypto processing (encryption/decryption), query processing, proxy delays, and the total delay. We use analytical evaluation and emulation. When emulation is used, each experiment is repeated a hundred times, and we report the average delays.

4.2.1 Major Steps

(a) We first develop an analytical model in which delays are predicted on the basis of various parameters that characterize delays of communication, cryptographic

Table 2 The configuration details

Server	OS	CPU	RAM	HD	Network	Server
Master cloud	Linux	2 vCPUs	4 GB	160 GB	400 Mb/s	XAMPP server
Slaves cloud	Linux	2 vCPUs	4 GB	160 GB	400 Mb/s	XAMPP server
Proxy server	OS X	2.4 GHz Intel Core i5	10 GB	500 GB	150 Mbps	Tomcat server 7.0.53

operations, query processing, and proxy delays. To do the evaluation using the developed analytical model, we need a set of experiments to measure the parameters that characterize the costs/delays. For instance, we measure the average delays for sending messages of various sizes. We then perform evaluation using a set of queries for which delays of user queries are predicted using the analytical method.

(b) We perform the evaluation using emulation in which we build the proxy, store the encrypted and plaintext relation in the master cloud, as well as store encrypted columns of the relation in various clouds for our method. The same queries are used as for the analytical model, and we measure delays for a set of experiments.

(c) Finally, we analyze the results.

4.2.2 Set of Queries

The queries used for evaluation are categorized into three groups as follows:

1. Select statements in which the Where clause contains one simple predicate.
2. Select statements in which the Where clause contains a conjunction of two predicates (connected by the AND Boolean operator).
3. Select statements in which the Where clause contains a conjunction of three predicates (connected by the AND Boolean operator).

4.3 Evaluation Using the Analytical Model

4.3.1 Analytical Model

Processing a query incurs the following costs/delays: communication delay, crypto (encryption/decryption) delay, query processing delay, and proxy delay.

Communication Delay

We model the communication delay of a message transfer by using the generally accepted Eq. (3):

$$D_s = a + bx \quad (3)$$

The delay is a linear function of a set-up overhead and the size of the message. As we are using Internet for communication, we are not able to determine the maximum size of packets, and hence we model communication on a message basis, as opposed to a more detailed modeling in which a message is sent in packets if it exceeds the maximum packet size. The communication delay is affected by the propagation delay, serialization, data protocol and latency, routing and switching latencies, and queuing and buffer management, and each of these factors has an impact on the total delay [24]. Since the messages are going over Internet, however, we have no control or real knowledge on the individual delay components and hence resort to a simple model represented by Eq. (3):

Crypto Delay

The crypto delay is a delay for encryption/decryption and is modeled by Eq. (4).

$$D_c = cn \tag{4}$$

Crypto delay is directly proportional to n, where n is the number of items to be encrypted/decrypted and c is the cost of encrypting/decrypting one item.

Query Processing Delay

Under the general term of query processing delays, we include delays for the basic SQL operations of Insert, Delete, Update, and Select.

1. **Insert**

   It depends on the number of tuples to be inserted, so the time complexity to insert into a database is $\mathcal{O}(n)$, where the n is the number of tuples.
2. **Delete/Update/Select**

   To find the tuples affected by the operation, the tuples need to be identified, and thus the delay depends on finding the tuples. Before we provide equations, we note that the delay to select/find tuples of a relation depends on whether there is a fast access data structure that can be exploited to find the desired tuples.

   (i) If there is no fast access method, all tuples are scanned, and hence the delay is directly proportional to the number of tuples in the relation, which is modeled by pn where n is the number of tuples in the relation and p is the delay to handle one tuple. If there is a fast access method, then the delay is proportional to $\log n$, where n is the number of tuples, and it is modeled as $p \log n$. Once the tuples are identified, they are processed as per delete, update, or select operation that we assume causes equivalent/same delays. However, this processing delay is directly proportional to the number of affected tuples. We need to note that the above simplifications are reasonable only under the assumption that the select queries are one variable only, that is, that they do not involve a join.

   (ii) As we deal with an evaluation in which we only have a single relation, we assume that the query does not specify a self-join. Furthermore, the equation also does not properly represent delays for an SQL query that has a group-by operator that would result in sorting of resulting tuples.

Since our experimental relation is relatively small-sized, we do not build indices explicitly, and we model the Delete/Update/Select by Eq. (5) under the assumption that there is no fast access method available and hence the processing delay is directly proportional to the number of tuples in the relation:
one tuple.

$$D_p = c + pn + psn \tag{5}$$

In Eq. (5), c is the overhead delay, n is the number of tuples in a relation, p is the delay to process one tuple, and s is the selectivity factor. The delay to process a query is thus modeled by overhead delay, c; delay of $p\,n$ for scanning all tuples of the relations for those tuples that satisfy the predicate, which could be a conjunction of simple terms; and delay of $s\ p\,n$ to handle the result.

Proxy Delays

The proxy is playing an important role as it performs several tasks that include encryption, decryption, query parsing, and key management. The proxy parses the query and rewrites it into another query or queries depending on to which slave sub-queries need to be sent. A number of sub-queries will be issued if there is a conjunction of simple terms. The time complexity for rewriting a query is modeled as $\mathcal{O}(1)$. Although there are a number of slaves, the number is small and fixed.

The proxy does both the encryption and the decryption, so the delay is already modeled within a separate section dealing with crypto delays. However, the key management is one of the proxy duties, so its delay needs to be included in the proxy overhead delay. The time complexity is equal to a search in an array by the element's position, which is $\mathcal{O}(1)$ because the keys are chosen based on their position in the array. On the other hand, the proxy must determine which encryption algorithm is needed to encrypt the query(s) value. Since we have a fixed number of the encryption algorithms, the delay to determine the encryption algorithm is also $\mathcal{O}(1)$.

The proxy also deals with the unions and intersects as more than one slave cloud may be queried in our method. The retrieved results of each slave cloud are stored in an array at the proxy. The proxy performs the union and the intersect upon the two arrays. In case there are more than two slave clouds accessed in processing a query, the result of intersect/union of two arrays will be stored on a temporary array where the intersect with a third array can be performed and so on. Thus, the delay for a union is $\mathcal{O}(n + n)$, and the intersect operation is $\mathcal{O}(m \log n)$ where m is the size of the first array and n is the size of the second array. Equation (6) is used to model the proxy's delay:

$$D_x = ax + by + c \tag{6}$$

In Eq. (6), the variable x is the number of elements in the first array, y represents the size of the second array, and a and b are constants representing the delay to process a tuple in array with the sizes x and y, respectively.

4.4 Delays Derived Using the Analytical Model

We are now ready to report on delays predicted by the analytical model. We explore two scenarios for our Secure Distributed method, one is serial and another is parallel. That is, if a user query results in more than one sub-query sent to the clouds in which columns are stored, we analyze two cases: (a) The sub-queries are issued in a serial fashion, one after another, and (b) sub-queries are issued in parallel, that is, concurrently. Once the results from sub-queries are retrieved, the proxy determines the intersection of tuples of the results. Based on the returned indices, this intersection is then sent to the master to get the query answer. We report on delays obtained by applying the analytical model on the queries described in Sect. 4.2.2. Recall that our proposed method has two variations, one serial and one parallel, that are compared to two base methods, one being Unsecure and one being Secure Centralized. As the name indicates, in the Unsecure method, there is no security, and hence the SQL queries are issued against the plaintext Databased (DB) table from which answers are derived in the usual manner. The Secure Centralized method is as described in [15] in which the whole DB is encrypted and stored in the cloud. Each column of a table is encrypted using an algorithm that supports the anticipated predicates. We calculate the total delays of each of these approaches using the analytical model while varying the selectivity to vary the number of retrieved tuples and the number of predicates.

Table 3 shows the total delays in milliseconds for queries 1, 2, and 3. As query 1 has only one predicate, there is no difference in delays between the parallel and serial versions of the Secure Distributed method. The delays for the Secure Distributed (serial and parallel) methods are higher than for the Secure Centralized method because in the centralized case only the master cloud is accessed, while the Secure Distributed method, in addition to accessing the master cloud, also accesses one slave cloud. However, the Secure Distributed method provides more security due to partitioning and distribution of the data. The selectivity factor of the predicate is varied from 0.2 to 1.0 in steps of 0.2. For each selectivity factor, there are component delays, namely, communication, query processing, crypto, and proxy, as is appropriate, these are also reported. Recall that in the Unsecure Centralized method, a query is processed in the cloud by scanning all tuples of the table. As the query result size is determined by the selectivity factor and the table is reporting delays per selectivity factor, there is no difference in delays for individual queries in this method, and hence, in Fig. 2a, we show the delays for the Unsecure Centralized method only once, that is, we do not show the delay for each individual query. Furthermore, the Unsecure method does not have the crypto component as the table is stored in the cloud in plaintext. Same statements apply for the Secure Centralized

Table 3 The total delays in milliseconds for queries 1, 2, and 3 using the analytical model

	Query 1					Query 2					Query 3				
S.F	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
Unsecure Centralized	130	160	186	211	237	134	160	186	211	237	134	160	186	211	237
Secure Centralized	339	569	798	1028	1258	339	569	798	1028	1258	339	569	798	1028	1258
Secure Distributed serial	445	674	904	1134	1363	536	845	1008	1244	1479	634	879	1108	1350	1585
Secure Distributed parallel	445	674	904	1134	1363	469	740	931	1163	1396	497	729	960	1111	1425

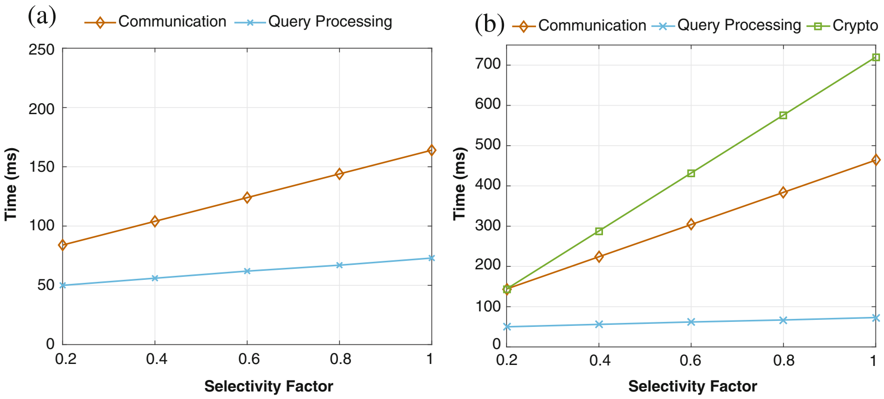


Fig. 2 The delays in milliseconds of (**a**) and (**b**) schemes for queries 1, 2, and 3. (**a**) Unsecure Centralized. (**b**) Secure Centralized

method, except that the method does have a crypto delay component. In that method, once the query is shipped from the proxy to the cloud, the query execution scans all tuples of the relation to identify the tuples that satisfy the predicate—hence there is no difference in delays in execution of queries in the cloud, and the query result size is determined by the selectivity factor—hence, in Fig. 2b we show delays only once and not for each individual query.

In Unsecure Centralized approach, unsurprisingly, the communication delay is higher than the query processing delay, and, of course, it has the smallest delays of all methods. The Secure Centralized method has higher delay than the Unsecure Centralized method, but the delay is smaller than those of the variants of the Secure Distributed method. As expected, for all methods, increase in the query complexities due to an increase in the number of predicates causes increase in delays. For the Secure Centralized method, in which the relation has encrypted columns in one cloud, there are communication delays to access the cloud, query processing delays, and crypto delays. For the Secure Distributed serial and parallel variants, there are four component delays: communication delays to access the master and the slave clouds, crypto delays, query processing delays on slaves, and proxy delays.

Figures 3, 4, and 5 show the component delays in milliseconds for queries 1, 2, and 3. Recall that a query i has i predicates. Delays for the parallel variant of the Secure Distributed method are less than those of the serial variant. It may also be observed that the crypto delays, although less than the communication delays, are significant. Further discussion of evaluation is offered below in Sect. 4.5.

4.4.1 Delays Derived Using Emulation

We measure the total delays while varying the selectivity factor and the number of predicates in the Where clause. We start the experiments with a query that has

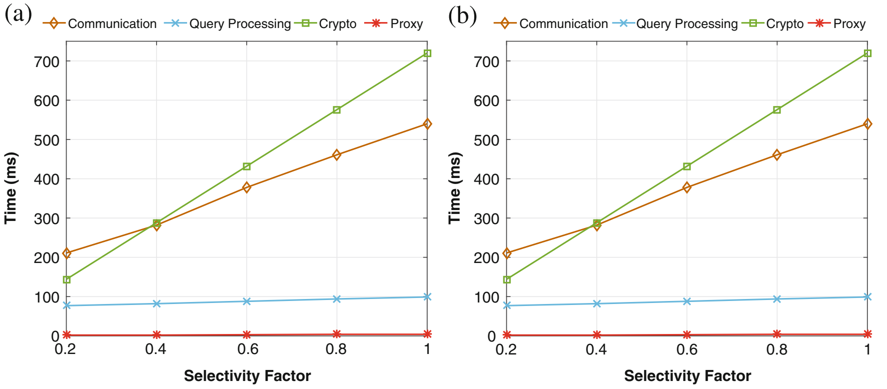


Fig. 3 The delays in milliseconds of (**a**) and (**b**) schemes for query 1. (**a**) Secure Distributed serial. (**b**) Secure Distributed parallel

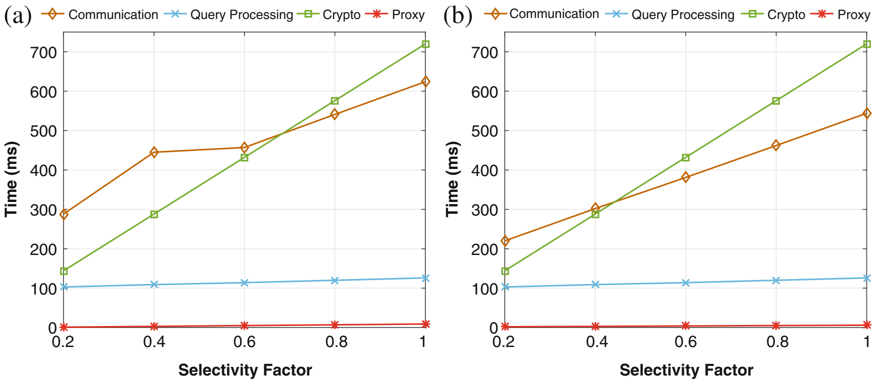


Fig. 4 The delays in milliseconds of (**a**) and (**b**) schemes for query 2. (**a**) Secure Distributed serial. (**b**) Secure Distributed parallel

a Where clause with one predicate and then we increase the number of predicates by one up to three predicates. We submit the queries a hundred times and report the averages. The experiments are applied for the Unsecure Centralized, Secure Centralized, and the serial variant of the Secure Distributed method. Table 4 shows the total delays in milliseconds for the three methods for queries 1, 2, and 3, respectively. Of course, the Unsecure method has shorter delays not only because there is no decryption process after the result is retrieved but also because the information is stored in clear text, which means that the tuples are smaller in size, resulting in less communication volume and hence delays.

Figure 6 shows the component delays in milliseconds for the three approaches for query 1, which has one predicate. The higher communication delay for the Secure Distributed method, in comparison to the Secure Centralized method, is due to the

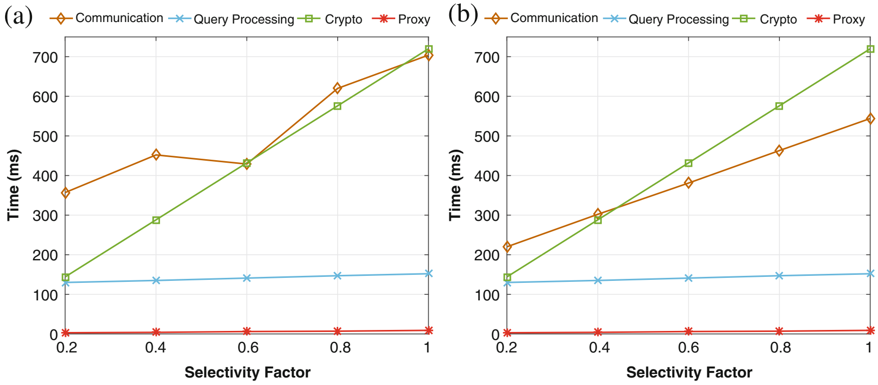


Fig. 5 The delays in milliseconds of (**a**) and (**b**) schemes for query 3. (**a**) Secure Distributed serial. (**b**) Secure Distributed parallel

communication delays of the extra round trip. It should be noted that the scale of y-axis values is not the same in Fig. 6a–c.

Figure 7 shows the delays for query 2 that has two predicates. It shows that the communication delay is much higher for the Secure Distributed method due to communication to process each predicate, which is done in a serial fashion. Figure 8 shows the component delays for query 3 with three predicates. The communication delay becomes the dominant factor for the Secure Distributed approach, increasing with the increase in the selectivity factor.

4.5 Result Analysis and Discussion

In the earlier section, we reported the component delays for each approach. In this section, we will compare the delays of each component of the analytical model with the corresponding delays of the emulation modeling across all queries. Figure 9 shows the communication delays, derived through the analytical modeling and emulation for all queries for Secure Distributed approach. In the Unsecure and Secure Centralized approaches, the communication delays are increased mostly due to the increases of the selectivity factors that produce larger results/relation that need to be communicated over the network. In our scheme, however, the communication is further increased as the number of predicates increases, as there is a sub-query issued for each predicate—hence increase in communications.

The differences in delays between the analytical model and the emulation model are higher in query 3 than in queries 1 and 2, and the trend increases with the increase in the size of data that needs to be processed and communicated, that is, with an increase in the selectivity factor. Clearly, analytical modeling

Table 4 The total delays in milliseconds for queries 1, 2, and 3 using emulation evaluation method

	Query 1					Query 2					Query 3				
S.F	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
Unsecure Centralized	130	115	143	152	171	153	160	185	200	210	155	175	207	273	220
Secure Centralized	451	585	804	983	1287	493	710	884	972	1582	504	660	775	1194	1441
Secure Distributed serial	777	1147	1492	1757	2073	976	1459	1707	2165	2223	1196	1729	1950	2291	2763

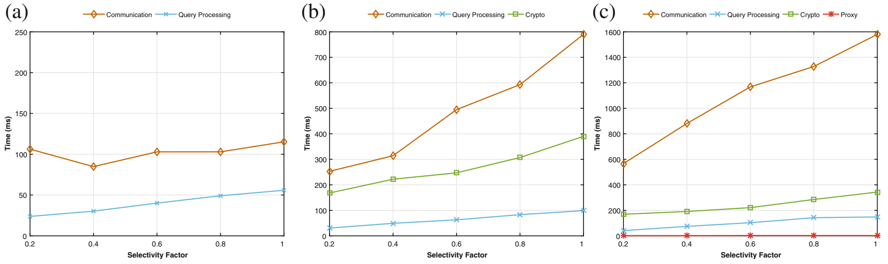


Fig. 6 The delays in milliseconds of (**a**), (**b**), and (**c**) schemes for query 1. (**a**) Unsecure Centralized. (**b**) Secure Centralized. (**c**) Secure Distributed serial

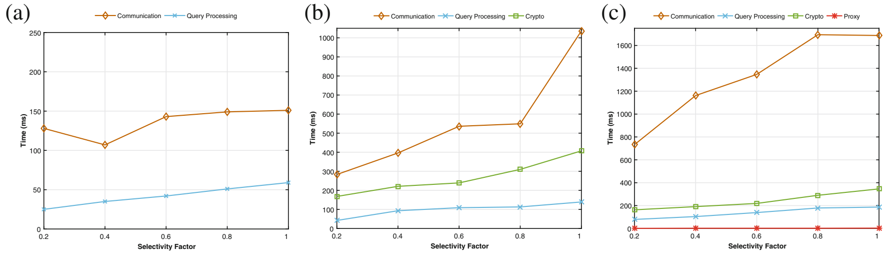


Fig. 7 The delays in milliseconds of (**a**), (**b**), and (**c**) schemes for query 2. (**a**) Unsecure Centralized. (**b**) Secure Centralized. (**c**) Secure Distributed serial

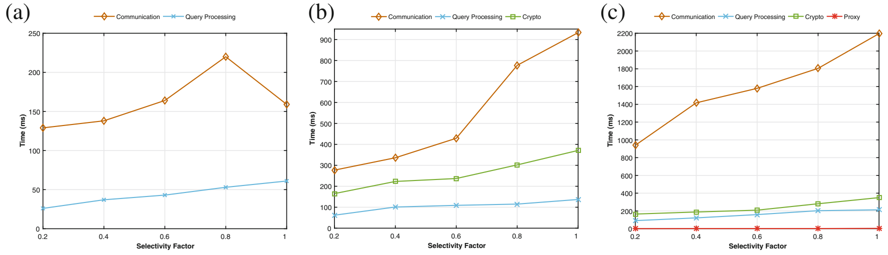


Fig. 8 The delays in milliseconds of (**a**), (**b**), and (**c**) schemes for query 3. (**a**) Unsecure Centralized. (**b**) Secure Centralized. (**c**) Secure Distributed serial

makes simplifying assumptions that are not necessarily reflected in real operations. However, the trends are correctly predicted in the analytical method.

Figure 10 shows the query processing delays, derived through the analytical modeling and emulation, for all queries for Secure Distributed approach. It can be observed that for small selectivity, delays through analytical modeling are higher than those obtained through emulation. However, for higher selectivity, delays obtained via emulation are much higher than those obtained by analytical modeling.

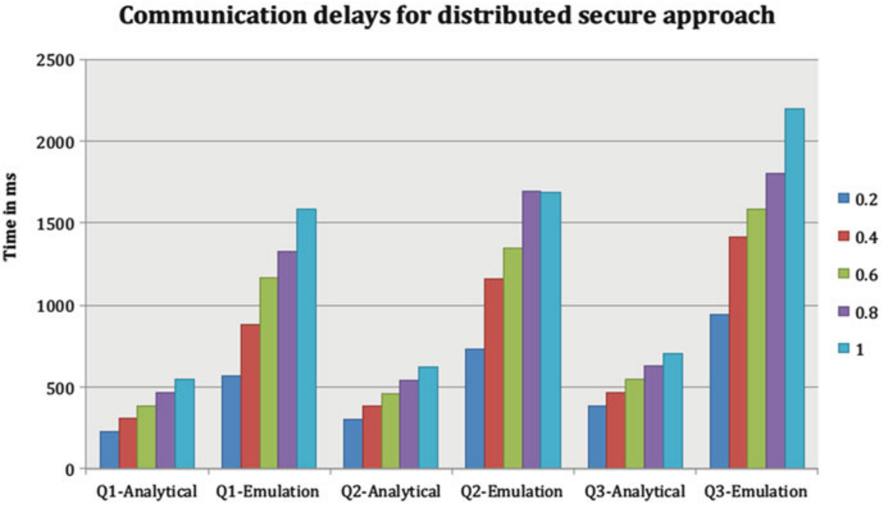


Fig. 9 Communication delays for the Secure Distributed Approach

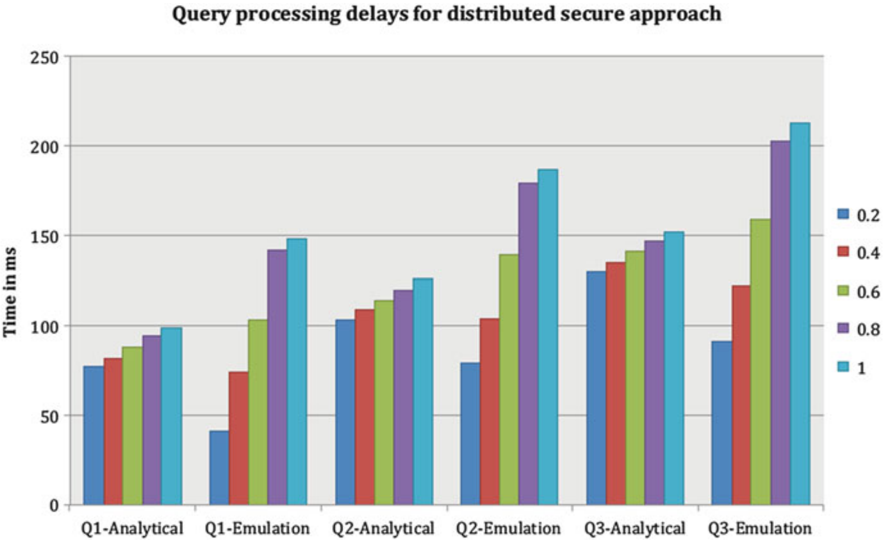


Fig. 10 Query processing delays for the Secure Distributed Approach

Figure 11 shows the crypto processing delays, derived through the analytical modeling and emulation, for all queries for the Secure Distributed approach. The crypto delays are directly proportional to the size of data to be encrypted or decrypted. Recall that the decryption process is needed only for the final result retrieved from the master cloud, and therefore, the measurement of the

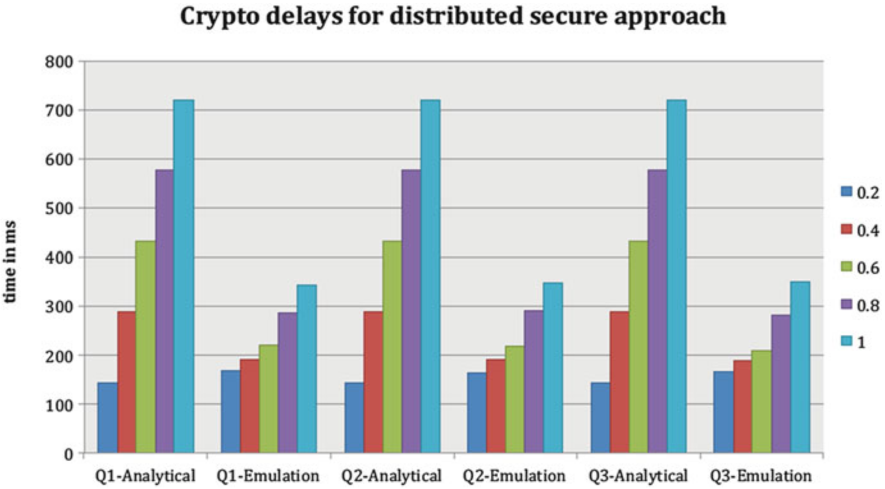


Fig. 11 Crypto delays for the Secure Distributed Approach

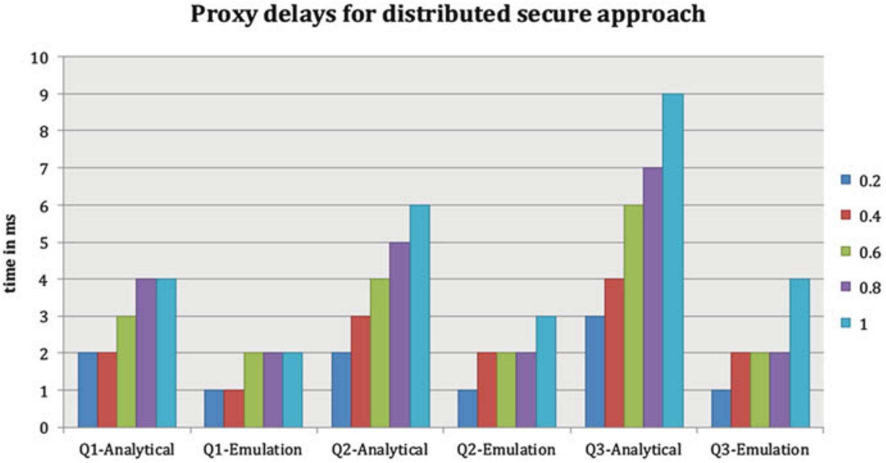


Fig. 12 Proxy delays for the Secure Distributed Approach

decryption operation depends on the selectivity factor. Thus, the delays obtained by the analytical model are the same for both the Secure Centralized and Secure Distributed approaches. The same applies for the delays derived through emulation.

Figure 12 shows the proxy processing delays, derived through the analytical modeling and emulation, for all queries for the Secure Distributed approach. Recall that the proxy delays include delays to perform the intersection of results returned from slave clouds and delays associated with transforming the user query to subqueries on clouds as well as the processing results obtained from the clouds. In general, the proxy delays in both models (i.e., analytical and emulation) are much less than the delays of the other components. The disparities in the delays between analytical and emulation approaches for the proxy delays are much lower than the disparities in the other components.

4.6 Discussion on Analytical Versus Emulation Modeling

We observe that the delays derived from analytical modeling, except crypto delays, are smaller than those derived from emulation. In the analytical model, we determine the component delays separately. Thus, the sum of an individual component delays forms the total delay of the analytical model. The analytical model is driven by first calibrating the models' parameters through measurements of average delays in communication, query processing, and decryption and then using such parameters in analytical modeling to predict the delays of the methods under consideration. When making the measurements, we use communication over the Internet and measure decryption and query processing delays that are performed on infrastructure provided by the cloud. Delays due to emulation are also obtained by using the cloud infrastructure for query processing and decryption, the Internet for communication, and thus, variability in both the analytical and the emulation approaches can be expected because both depend on when measurements are being done and the load on the environment when measurements are being done. Although this variability in performance is obvious for Internet delays, it also applies to measurement of the other components as the processing is performed in the cloud. Delays in cloud infrastructure are also variable as they depend on the load, resources allocated by the cloud provider, and the cloud's quality of software that is used to measure the load and to allocate resources to handle it.

To conclude this section, although analytical modeling underestimated delays in comparison to emulation, it is useful as it showed the general trend in delays and thus is well suited to provide guidelines on the general trend when variables, such as selectivity or type of issued queries, are varied.

5 Conclusion

Cloud computing technology is attractive for storing and managing data. However, concerns over confidentiality concerning storing sensitive data prevent many public and commercial organizations from moving to the cloud. A number of researchers have attempted to provide solutions to the security concerns associated with outsourcing storage. The aim of this research was to investigate how to prevent untrustworthy cloud providers from obtaining sensitive data. We proposed a combination of encryption algorithms in [15] and obfuscation by distributing data among different clouds for confidentiality of storing data in the cloud. Specific encryption algorithms provide the user with strong confidentiality but when querying of encrypted data is required than specific encryption algorithms, which support querying of encrypted data, may have to be used even though they may not be strong enough for certain type of attacks. However such attacks may be thwarted by through obfuscation by distribution of data. We built a prototype for our proposed method to demonstrate its feasibility, determine overhead delays, and compare it to base methods reported in the literature.

We used two modeling techniques to determine the delays of our proposed method and the base methods used for comparison. One method was based on an analytical model, while the other method used emulation using a prototype. We calibrated our analytical models' parameters by measurements. When we compared delays derived through analytical modeling to those derived through emulation, we observed that in most cases the analytical model underestimated the delays. Thus, the analytical model can be used to analyze the behavior of the system regarding the trends on delays, but it cannot be used to predict such delays accurately. Of course, the same can be applied for any method that is predicting delays when the load on the system is not taken into account, the load in our case being the load that affects the communication over the Internet and the load that affects the processing delays on a cloud infrastructure.

Acknowledgments This work is partially supported by Aljouf University represented by the Saudi Arabian Cultural Bureau in Canada. The authors would like to thank the anonymous reviewers for their constructive comments.

References

1. P. Mell, T. Grance, and T. Grance, "The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology," 2011.
2. A. Amjad, P. Bodorik, and S. Sampalli, "Improving database security in cloud computing by fragmentation of data," in *2017 International Conference on Computer and Applications (ICCA)*. IEEE, Sept 2017, pp. 43–49.
3. H. Hacig, "Query Optimization in Encrypted Database," pp. 43–55, 2005.
4. N. Anciaux, M. Benzine, L. Bouganim, P. Pucheral, D. Shasha, and I. Rocquencourt, "GhostDB : Querying Visible and Hidden Data Without Leaks," 2007.
5. A. Hudic, S. Islam, P. Kieseberg, S. Rennert, and E. R. Weippl, "Data confidentiality using fragmentation in cloud computing," *International Journal of Pervasive Computing and Communications*, vol. 9, no. 1, pp. 37–51, 2013. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84878829696&partnerID=tZOtx3y1
6. H. Hacig and C. Li, "Executing SQL over Encrypted Data in the Database-Service-Provider Model," vol. 7, 2002.
7. B. Hore, S. Mehrotra, M. Canim, and M. Kantarcioglu, "Secure multidimensional range queries over outsourced data," *The VLDB Journal*, vol. 21, no. 3, pp. 333–358, Aug 2011. [Online]. Available: http://link.springer.com/10.1007/s00778-011-0245-7
8. B. Hore, S. Mehrotra, and G. Tsudik, "A privacy-preserving index for range queries," pp. 720–731, Aug 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1316689.1316752
9. L. Bouganim and P. Pucheral, "Chip-Secured Data Access : Confidential Data on Untrusted Servers," 2002.

10. S. Y. Ko and K. Jeon, "The HybrEx Model for Confidentiality and Privacy in Cloud Computing," 2011.
11. K. Zhang, X. Zhou, Y. Chen, and X. Wang, "Sedic : Privacy-Aware Data Intensive Computing on Hybrid Clouds Categories and Subject Descriptors," pp. 515–525, 2011.
12. Z. Zhou, H. Zhang, X. Du, P. Li, and X. Yu, "Prometheus : Privacy-Aware Data Retrieval on Hybrid Cloud," pp. 2643–2651, 2013.
13. C. Zhang, E.-c. Chang, and R. H. C. Yap, "Tagged-MapReduce : A General Framework for Secure Computing with Mixed-Sensitivity Data on Hybrid Clouds," pp. 31–40, 2014.
14. K. Y. Oktay and S. Mehrotra, "SEMROD : Secure and Efficient MapReduce Over HybriD Clouds The University of Texas at Dallas," pp. 153–166, 2015.
15. R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan, "CryptDB : Protecting Confidentiality with Encrypted Query Processing," pp. 85–100, 2012.
16. E.-O. Blass, G. Noubir, and T. D. Vo-Huu, "Epic: Efficient privacy-preserving counting for mapreduce," Cryptology ePrint Archive, Report 2012/452, 2012, http://eprint.iacr.org/2012/452.
17. J. J. Stephen, S. Savvides, R. Seidel, and P. Eugster, "Practical confidentiality preserving big data analysis," in *6th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 14)*. Philadelphia, PA: USENIX Association, Jun. 2014. [Online]. Available: https://www.usenix.org/conference/hotcloud14/workshop-program/presentation/stephen
18. T. Mayberry, E.-o. Blass, and A. H. Chan, "PIRMAP : Efficient Private Information Retrieval for MapReduce," pp. 371–385, 2013.
19. E.-o. Blass, R. D. Pietro, R. Molva, and M. Onen, "PRISM — Privacy-Preserving Search in MapReduce," pp. 180–200, 2012.
20. S. D. Tetali and T. Millstein, "MrCrypt : Static Analysis for Secure Cloud Computations," pp. 271–286, 2013.
21. D. Liu, S. Wang, and C. I. C. T. Centre, "Programmable Order-Preserving Secure Index for Encrypted Database Query," *2012 IEEE Fifth International Conference on Cloud Computing*, pp. 502–509, Jun. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6253544
22. D. Xiaodong, S. David, and W. Adrian, "Practical Techniques for Searches on Encrypted Data."
23. N. Singhal and J. P. S. Raina, "Comparative Analysis of AES and RC4 Algorithms for Better Utilization," pp. 177–181, 2011.
24. W. Stallings, *Cryptography and network security: Principles and practice.* Upper Saddle River, N.J. : Prentice Hall, 1999.
25. P. Samarati and I. C. Society, "Protecting Respondents' Identities in Microdata Release," vol. 13, no. 6, pp. 1010–1027, 2001.
26. J. Daemen, *The design of Rijndael : AES - the advanced encryption standard with 17 tables*. Berlin [u.a.]: Springer, 2002.
27. J. Blomer, "Fault Based Cryptanalysis of the Advanced Encryption Standard (AES)," *Lecture notes in computer science.*, no. 2742, pp. 162 – 181, 2003.
28. E. M. Mohamed, "Enhanced Data Security Model for Cloud Computing," pp. 12–17, 2012.
29. A. Arasu, S. Blanas, K. Eguro, M. Joglekar, R. Kaushik, D. Kossmann, R. Ramamurthy, P. Upadhyaya, and R. Venkatesan, "Secure database-as-a-service with Cipherbase," *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*, p. 1033, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2463676.2467797
30. G. Nalinipriya and R. Aswin Kumar, "Extensive medical data storage with prominent symmetric algorithms on cloud - A protected framework," *International Conference on Smart Structures and Systems - Icsss'13*, pp. 171–177, Mar. 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6623021
31. "Rackspace: The Leader in Hybrid Cloud." [Online]. Available: http://www.rackspace.com/

CPU/GPU Hybrid Detection for Malware Signatures for Battery-Powered Devices Using OpenCL

Radu Velea, Ştefan Drăgan, and Florina Gurzău

1 Introduction

Malware is the designated term for any malicious software that disrupts the normal workflow of a computer application or system. Malware can take the form of executable code that passes itself as legitimate in order to compromise a system. A compromised system may become vulnerable to additional cyber-attacks and this in turn can lead to information loss or theft, denial of service and other undesired consequences. Malware scope can range from single users to organizations or public infrastructure.

The definition of the term is broad and it can be used to refer to any kind of software that has a negative impact on user experience or damages assets. Common types of malware include ransomware, viruses, spyware or other types of computer viruses. Versions of these entities can be found in just about every known form

This is an extended version of a preliminary conference paper that was accepted and presented in ICCA 2017 [1].

R. Velea (✉)
Department of Computer Engineering, Technical Military Academy of Bucharest, Bucharest, Romania

Bitdefender, Bucharest, Romania
e-mail: rvelea@bitdefender.com; radu.velea@mta.ro

Ş. Drăgan
Bitdefender, Bucharest, Romania
e-mail: sdragan@bitdefender.com

F. Gurzău
Department of Computer Engineering, Technical Military Academy of Bucharest, Bucharest, Romania
e-mail: florina.gurzau@mta.ro

J. M. Alja’am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_9

factor: from embedded devices to supercomputers. A system is usually contaminated through an infected file. The infected file can take advantage of a vulnerability inside a legitimate application and execute malicious code. Infections can come via hard drives, USB sticks, and optical storage devices or from the network. In the age of the Internet, where most devices are interconnected, unchecked infections can spread rapidly. To mitigate this risk, multi-layered defense systems have been implemented to protect endpoints and networks. Typical applications that combat malware consist of antivirus software, firewalls, network intrusion detection and prevention systems (NIDPSs).

In the current environment, the attack surface for malicious applications is very large. Lack of transparency and reaction from vendors can result in security incidents that go unnoticed for weeks or even months. Detection and prevention tools usually rely on matching suspected files against a database of known threats. The security of a system can greatly depend on how often its signature database is updated with the latest discovered threats. While organizations tend to have a process in place for protection against attacks, the average user has to rely on his or her provider. Lack of technical knowledge causes users to fall prey to viruses that have been known to exist for years or decades. Even if their service provider supplies regular security updates, users may be unwilling to adopt them due to performance considerations (fear it may slow down their device, take too much disk space, bandwidth, etc.) or sheer ignorance [2].

Customer and enterprise security solutions are required to process ever-increasing amounts of data. Normal workloads include scanning files on the disk, examining process memory, validating user input or filtering network traffic. The growing complexity of new cyber-threats means that detection and protection tasks need to consume more resources in order to be efficient.

The context described above drives security companies to develop performance-efficient solutions to cope with the ever-growing amount of malicious content their customers are exposed to. In the following sections, we will describe what makes malware detection troublesome from a performance point of view and underline the hotspots antimalware tools have. We will then propose a new method to speed up detection by combining the compute capabilities found in consumer desktop systems and laptops.

Detection of malicious code can be done statically or at runtime. Static analysis requires a set of pattern matching operations that have to determine if a blob of data resembles any known malware. Researchers try to populate databases with malware signatures that will then be used to scan files. To counter this, malware writers go to extreme lengths to obfuscate their code and bypass any known filters. A practical way to detect a malware instance is to generate a footprint by performing semantic rather than syntactic analysis of the code [3]. Another frequent challenge is dealing with zero-day vulnerabilities and self-mutating malware. These kind of attacks can be mitigated through runtime analysis (executing the code in a contained environment and observing its behavior) or through the implementation of machine learning algorithms. Applying such techniques in real time can take up a significant amount of resources and generate false positives. An option would be to perform these investigations in a controlled environment and then generate static signatures that can detect the new threat and its derivatives in the wild.

Once the signature has been generated, it can be used by compatible tools to detect that type of malware. Detection could happen on a variety of devices: embedded systems, mobile phones, desktops, cloud infrastructure, etc. For example a vulnerability in a web browser could affect all systems that access the Internet through it. Protection systems are thus deployed in different forms according to the device's available resources and its security needs. A network sensor that runs an NIDPS would be responsible for deep packet inspection. Studies have shown that a common performance bottleneck during this process is the necessity to perform string matching [4, 5]. The overhead of pattern matching can cause degradation of network performance, while relaxing the rules could allow threats to go undetected. A host-oriented software such as an antivirus is required to perform regular scans in order to ensure the integrity of the system. If these regular scans take too long or strain the machine's resources, the user might opt to perform them at longer intervals or skip them altogether.

It is therefore important that the patterns matching process required for malware detection be performed in an efficient manner and that it takes advantage of all the computing resources available on the host device.

2 Related Work

2.1 Detection Through String Matching

Static malware detection comes down to string searches—finding known blocks of malicious code inside data on the system or network. String matching algorithms based on Aho-Corasick [6] and Boyer-Moore [7] have been adapted for this task. They are used by commercial software as well as open source projects like Snort, Suricata or ClamAV. The logic behind these implementations is to perform the minimum number of byte comparisons on the smallest set of data possible without compromising the accuracy of the search. The theoretical details behind these algorithms are out of the scope of this work.

Hashing techniques can accelerate pattern matching algorithms and can potentially detect viruses encrypted with simple functions such as ADD and XOR [8].

Efficient hash functions that provide few collisions over a set of characters could be used instead of conventional string matching operations. This approximate solution to detect threats can be useful in scenarios where the set of input characters is reduced—for example, when scanning scripts, markup language or human-readable text.

2.2 Parallel Implementations

A survey performed by [9] estimated that as much as 75% of CPU time is spent performing pattern matching in NIDPSs. High traffic throughput has motivated researchers to look for alternative ways to offload scan tasks that would normally run on the CPU. The GPU is an ideal candidate for this assignment because of its SIMD architecture and high level of parallelism. Experiments with Snort [10] have concluded that GPU string matching is efficient, but that performance can deteriorate if memory transfers are not handled accordingly. This can make real-time detection problematic if the GPU is used to scan packets that are few and far between or small, individual files on the disk.

Changes to the algorithms that run on the GPU, focus on optimizing memory accesses [11] and exploiting the large number of available compute units [12]. Favored approaches include the compression [13] of the state machine for automatons and removing the failed transactions [14] (the current thread will exit after a character mismatch rather than try to continue from another valid state). For algorithms based on lookup tables, an optimization would be to use hashed prefixes [15] to skip as many characters as possible from the benign input.

Some works have proposed hybrid implementations that use OpenMP together with CUDA to perform string matching [16, 17]. Memory limitations negatively impacted the number of signatures that could be searched in the string, but the solutions provided significant speedups over the serial versions. To solve some of the drawbacks caused by transferring data back and forth between the GPU and CPU, developers have looked for alternative solutions that can perform opportunistic load balancing [18] or shallow searches. GPU hardware vendors have advertised new designs that promise to solve this problem by providing a unified memory model [19]. Some of the devices available on the market that share memory between CPU and GPU are mobile phones and other small form factors (ultrabooks, laptops, chromebooks) with integrated graphics.

Software frameworks used for GPU programming are Compute Unified Device Architecture (CUDA) and OpenCL. CUDA is the older and more popular technology. It is designed to run on NVidia hardware and besides graphics, it is used for high-performance computing in physics, medical imaging, distributed computing and other GPGPU-related work. OpenCL is an open standard maintained by a consortium of hardware and software vendors. It is designed to run on a greater variety of hardware and has both proprietary and open source implementations. OpenCL is a flexible API and can run on multicore CPUs and better map itself to low-end platforms. CUDA and OpenCL have similar memory and programming models. The solution described in this paper has been implemented in OpenCL. The choice of OpenCL over CUDA is motivated by the fact that OpenCL's role is to enable parallel programs to run across heterogeneous hardware and, as a result, is more widely available among users (NVidia graphics can run OpenCL applications).

The current paper explores the opportunity of using the GPU to offload compute-intensive tasks that usually take a lot of the CPU time. The case studies described in this work target battery-powered devices, such as laptops or ultrabooks, that implement some form of security software. The goal is to use the extra computing power of the GPU to improve execution times and reduce overall power consumption for system scans or other antimalware processes.

3 Implementation

3.1 Exact Pattern Matching

The current section describes our proof of concept for parallelizing malware detection across heterogeneous hardware. Our string matching implementation is based on a variation of Boyer–Moore–Horspool [20] algorithm. The algorithm was designed to search for malware signatures inside files located on the drive. We make the assumption that for the most part our searches will not result in any detection, regardless of the signature database size or file system. This detail will be used to provide an additional speedup during the scanning phase.

The static fingerprint of a piece of malware is defined as a set of instruction blocks that make it stand out from other conventional pieces of software. Malware signatures are available online and are updated regularly when new infections are discovered. These instruction blocks are the patterns we have to identify among the scanned content. Before building the lookup table we load all the signatures into memory and sort them using the first 32 bytes of their binary code as key. The resulting sorted structure will be used later to search for exact matches.

In the preprocessing stage, the lookup table is built with integer (4-byte) values rather than individual bytes. The integer values represent a hash of the last key bytes. Offsets or skip distances between input bytes are computed via hash equality and not byte equality. The hash function is not injective and occasional collisions will occur between signatures. This means that once a match is detected there is a chance it could be a false positive (a byte sequence that just happens to have the same hash as a malicious pattern). In order to mitigate this, the algorithm performs a binary search into the sorted key structure and checks for an exact match. This process is compute-intensive but it is only expected to be executed in exceptional cases. If any malware signatures are detected, they will be reported to the upper levels of the application and dealt with accordingly.

3.2 Approximate Pattern Matching

The method described above has some limitations: it requires the scan engine to load the signatures in their fullest form in order to perform an exact match. For simple, 32-bit hashes used for malware signature lookup we can expect to have regular collisions that have to be solved down the line by performing byte-to-byte comparisons. Given the nature of the algorithm we can expect to have a match for one out of one million bytes, regardless of the nature of the input. This means that we will have to perform an extra check for roughly 1 MB of scanned content. While this does not influence the scan speed significantly, it does impact memory consumption, as the large signature database has to be carried around for these extra checks. This makes deployment problematic for low-end devices because it ties down an amount of memory proportional to the size of the signature database.

To solve this problem, the scan engine would have to fully rely on hashes for matching incoming traffic against the malware database. This procedure is not 100% accurate, but can provide a solution that can satisfy the current needs within a reasonable degree of certainty. In this case, the scan engine would no longer perform byte-level comparisons to determine if a match is a false positive or not, instead relying on a combination of hashes and to determine the final result. Given the a token size of 256 bytes for a malware signature, we can look for hash functions that have a good spread and are easy to compute.

A solution would be to use the same hash function at different offsets. The malware signature would be preprocessed and the hash values would be stored in the process's memory instead of the actual contents. This method would allow our algorithm to reuse the latter computation at different stages of the detection process and would significantly reduce the memory footprint.

$$\begin{aligned} f(0) &= hash\,(offset\;0, 255) \\ f(1) &= hash\,(offset\;1, 256) \end{aligned} \tag{1}$$

If the computed values for $f(0)$ and $f(1)$ match the ones loaded for a particular signature, we can safely claim that we have match. To further reduce the possibility of having a false positive, multiple levels of the function f can be used. The mathematical chance that a signature matches benign content based on this mechanism still remains, but a careful analysis of the hash function and the type of input it is used on (e.g., human readable text, scripts and binary code) can be used to reduce it to almost 0. A further advantage in speed for serial implementations can be achieved if the $f(n+1)$ depends on $f(n)$. This dependency can help the algorithm's speed, but in turn adds a constraint that can hinder the parallelization process.

An interesting candidate for a fast hash function has been identified in CLHASH [21]. CLHASH uses the carry-less multiplication instruction CLMUL, available on x86 architectures to compute a non-cryptographic hash. The structure of the algorithm is based on performing multiple carry-less multiplications on chunks of bytes obtained by XOR-ing the input data with a randomly generated key, and then accumulating the result. The mechanism is simple enough to allow modifications—we will use some of the elements present in CLHASH to create a candidate-function that could be used on both the CPU and GPU. The original code of CLHASH was based on the *pclmulqdq* instruction from the expanded instruction set of x86-64 architectures. This instruction performs a carry-less multiplication on two 64-bit integer values and stores the result in a 128-bit register. This instruction provides significant speedup for the function on the CPU side.

3.3 OpenCL Parallelization for Exact Pattern Matching

The scan process involves parsing a large amount of input data, one byte at a time and identifying possible matches, as described above. Our OpenCL parallelization efforts focused on offloading this part of the code to the GPU. After the lookup-search is done in parallel, the results are transferred to the CPU for the binary search to complete the match and take necessary actions. This process is done one file at a time. Some of the related work presented [15] in the previous section suggests concatenating a significant amount of data before sending it to the GPU in order to minimize the penalty incurred from frequent memory transfers. While this approach may seem sound, it is not practical for most real-life scenarios. A user may decide to incrementally scan his hard drive, a few files at a time, or may simply not possess the available resources to load large amounts of data (GBs) into memory. Files most susceptible to infection are generally small in size [22]. For this reason, the focus of our implementation is to achieve equal or better speedups when using small amounts of data.

The most straightforward approach is to split the amount of data evenly across all available GPU threads. Each thread would receive a chunk of bytes and will have to report which of them are valid offsets for future analysis. To reduce the amount of computation on the CPU we first attempted to compute the candidate key for the next-stage binary search. Experimental results showed the penalty for repeatedly accessing GPU global memory to be significant. To reduce the number of memory accesses we also reduced the amount of computation and only outputted a corresponding bit value for each processed byte. Input data would be padded to 64 bytes and each GPU work item would be responsible to compute the output for a fixed chunk of 64 bytes. The output bits would be added to a 64-bit unsigned long mask and copied back to CPU memory. Each thread would only have to access the global memory that contained input bytes and lookup tables (which would only be transferred to the GPU once—after the signature preprocessing stage). The CPU would then iterate through the bitmasks received from the GPU and process any nonzero values. Boyer-Moore table lookups would ensure that each GPU thread will actually process less than 64 bytes, as most of them are expected to be skipped. Further parallelization can be done on the CPU side by using multithread libraries such as Pthread. Each Pthread would have a corresponding OpenCL context and handle its own content. This would allow for multiple files to be scanned in parallel, but would duplicate the amount of memory required on the GPU side, as lookup tables would not be shareable across contexts (Fig. 1).

The function responsible for preprocessing or filtering the input buffer should be chosen carefully. Some types of operations are known to cause significant performance penalties when used on the GPU (e.g., branching instructions).

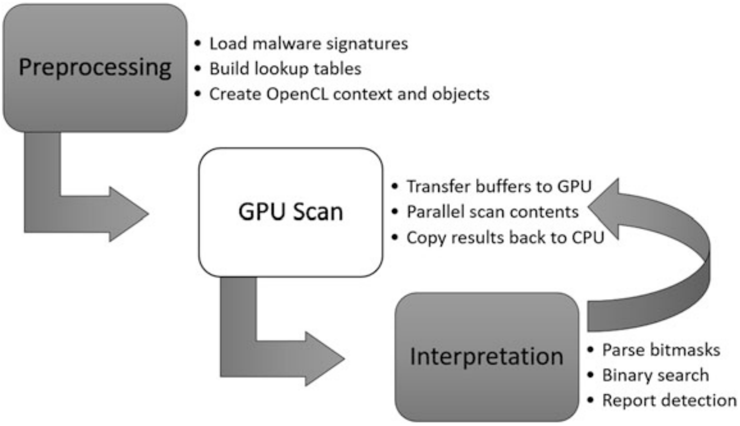


Fig. 1 Malware scan workflow

The current work presents a solution that scans one file at a time and uses on a single CPU thread with a single OpenCL context in which multiple work items are executed.

4 Results

4.1 Test Setup

The implementation was tested on a 64-bit ultrabook with Intel® CoreTMi7-6600U with integrated HD Graphics 520. This form factor can run Windows operating systems as well as Linux-based distributions, such as Ubuntu and ChromeOS, or Android. The integrated GPU has 24 compute units clocked at 1050 MHz. The machine does not have dedicated graphics memory, but instead uses a part of the main memory. Level 3 cache is shared between CPU and GPU. This setup is ideal for testing the performance of our hybrid detection framework. Similar devices are available on the market and used for business or leisure.

The malware database used for the tests contained approximately 20,000 signatures. The test files consisted of Windows system files, Linux root file system, and randomly generated input, along with some selected malware samples. In case a file was too large, a fixed-sized buffer was created in order to scan only X amount of data at a time.

4.2 Performance

Performance tests were categorized into two groups. The first group involved repeated scans using variable buffer sizes. This test will determine the speedup between the hybrid implementation and the CPU-only one. The sizes of the scanned files will range from a couple of bytes to several GBs in size. This method of evaluation will help us find the ideal buffer size that provides the best performance on our device.

Figure 2 shows OpenCL is not very efficient in scanning small buffers. Upon closer examination it was found that for a 1 KB buffer only 2% of computation time was spent on the GPU (either executing code or performing memory transfers). The rest of the time (about 0.3 ms) was spent in OpenCL library calls: sending commands to the execution queue, scheduling, waiting for other events, etc. To reduce part of this penalty memory transfers were performed by mapping GPU buffers into host address space and performing read and write (memcpy) operations on the CPU. As the size of the scanned buffer grows, the hybrid performance improves compared with the CPU-only. The point where hybrid performance surpasses the CPU is around a 4 MB buffer (the test machine has a 4 MB L3 SmartCache [23]). In Fig. 3 (as the buffer grows), GPU-time reaches around 96% of the total hybrid computation and speedup values increase from 1.12x to 2.57x in favor of the GPU+CPU solution.

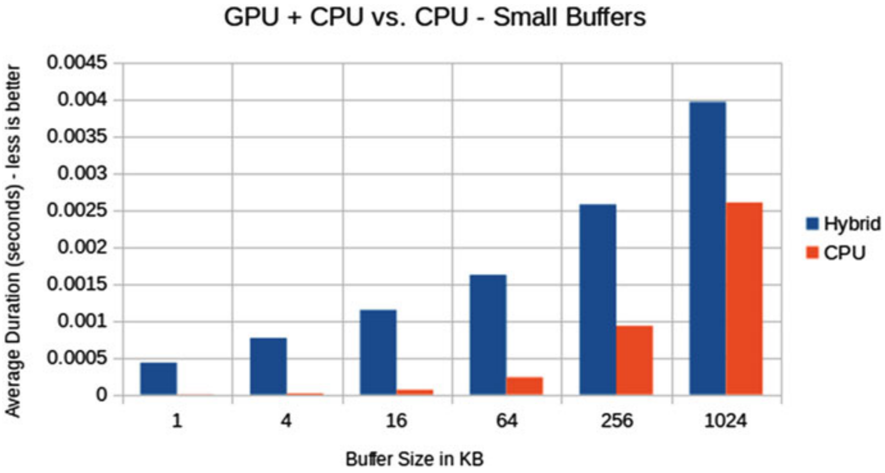


Fig. 2 Hybrid pattern matching performance on small buffers

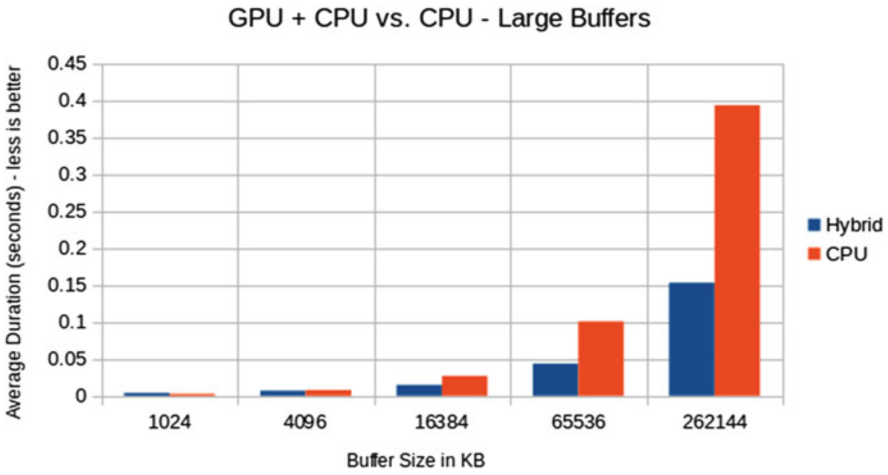


Fig. 3 Hybrid pattern matching performance on large buffers

Table 1 Test summary

	Hybrid	CPU-Only
Duration (s)	22.4	40.7
Avg. CPU Power (W)	4.68	6.5
Avg. GPU Power (W)	4.23	0.1
Total Energy (W * s)	199.58	268.62

The second group of tests focused on power consumption and overall efficiency. Tools like GPU-Z[1] and Intel® Power Gadget[2] were used to measure the power consumption and other metrics while scanning. We set the buffer size to 16 MB and measured the power consumption of the CPU and GPU:

The hybrid implementation is almost twice as fast and consumes 25% less power while running the benchmark (Table 1):

The test was performed while the device's power plan was set to high performance: CPU frequency was 3200 MHz for the duration of the test.

4.3 Experiments with Other Hash Functions

To create a more complex hash function, suitable for both the GPU kernel and the CPU, we experimented with some elements from CLHASH: we selected a scenario in which the scanned input consists of a stream of human-readable text and we generated hashes from the malware database to be used instead of byte-per-byte comparisons. For each 256-byte signature we would create two 64-bit hashes at

[1] https://www.techpowerup.com/gpuz/

[2] https://software.intel.com/en-us/articles/intel-power-gadget-20

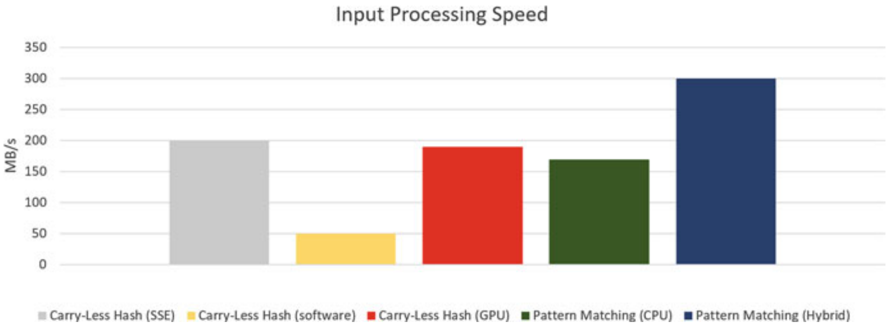


Fig. 4 Comparison of power consumption between hybrid and CPU-only

64-byte intervals and store them sorted in the memory. We would then process the input text and perform 64-bit lookups in order to determine if we have a match or not. CLMUL instructions are not available from OpenCL on the GPU. To mitigate this disadvantage, we created software versions of the hash functions and compared the results with the previous detection mechanism and the SSE[3] version of the hash function.

Figure 4 shows that in spite of the parallelization efforts, the carry-less hash function performs poorly without dedicated hardware support. Nevertheless, the performance is still a good 10% better than the best CPU-only pattern matching scheme. With future dedicated instructions that can perform carry-less multiplication on the GPU, there is the potential of achieving better results in terms of speed.

From a memory point of view, this scheme brings a significant reduction in the size of the memory used by the scan engine. With the original pattern matching framework, the process would have to store 256 bytes plus the size of the lookup table key for each of the malware signatures. Using an approach based exclusively on hashes can bring down memory by almost 75%.

5 Conclusion

The results presented in Fig. 5 suggest the hybrid solution offers a significant advantage in power and performance over a CPU-only implementation. However, these advantages disappear if the application has to scan small files (less than 1 MB in size). Based on these experimental results we could introduce a logic inside the application to take different code paths according to the amount of data available. For the selected test platform, the impact on battery power seems to favor this hybrid approach. If the application has to handle large amounts of data, a bigger internal GPU scan buffer would provide incremental benefits.

[3] Streaming SIMD Extensions.

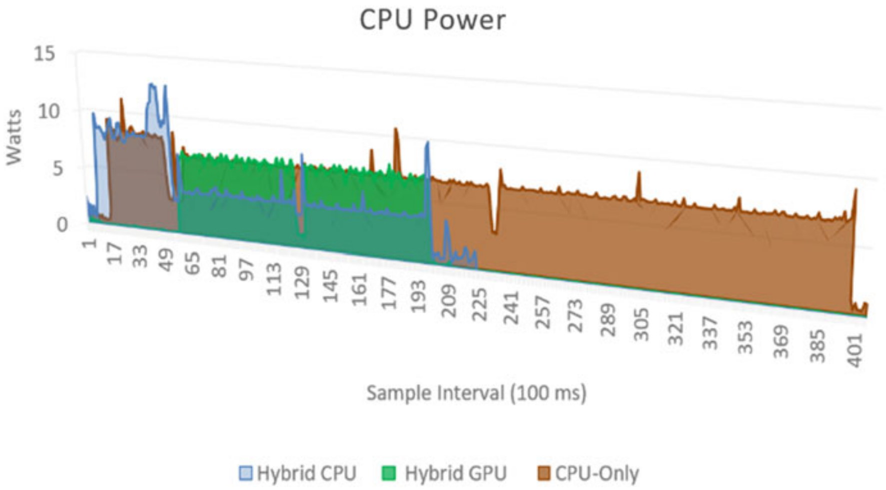


Fig. 5 Speed processing speed for hash-based implementations

Contextual scenarios can benefit more from frameworks that use a combination of hash values rather than classical pattern matching. The parallelization of such frameworks might not yield significant performance improvements compared to an optimized CPU-only version, but there is a promise for future improvement as hardware and software evolve.

The current proof of concept represents a step closer toward the creation of new security solutions that harness all the existing computing resources available on a machine. The concept of "security through total computing" looks to promote the implementation of heterogeneous software products that enhance the level of security of the average user, without compromising user experience.

Our conclusion is that efficient usage of an integrated GPU can speed up string matching operations for antimalware software on battery-based devices. The unified memory model provided by the test hardware reduced the overhead incurred by repeated memory transfers between CPU and GPU, but also created a penalty for multiple accesses of GPU global memory inside the kernel code and made usage of local memory impractical. OpenCL provides a versatile framework for offloading intensive computation performed in network and host intrusion detection systems in environments that have, otherwise, limited resources. Future work will include deployment and measurement on platforms that have dedicated GPUs and further comparisons with CUDA-based implementations and other string matching algorithms.

References

1. R. Velea and Ş. Drăgan, "CPU/GPU Hybrid Detection for Malware Signatures," in *Computer and Applications (ICCA), 2017 International Conference on*, Dubai, 2017.
2. Zhang-Kennedy, S. C. Leah and R. Biddle, "Stop clicking on "update later": Persuading users they need up-to-date antivirus protection," in *International Conference on Persuasive Technology*, 2014.
3. M. Christodorescu, S. Jha, S. A. Seshia, D. Song and R. E. Bryant, "Semantics-aware malware detection," in *Security and Privacy, 2005 IEEE Symposium on*, 2005.
4. K. Salah and A. Kahtani, "Performance evaluation comparison of Snort NIDS under Linux and Windows Server," *Journal of Network and Computer Applications,* vol. 33, no. 1, pp. 6–15, 2010.
5. P.-C. Lin, Y.-D. Lin, Y.-C. Lai and T.-H. Lee, "Using string matching for deep packet inspection," *Computer,* vol. 41, no. 4, 2008.
6. A. V. Aho and M. J. Corasick, "Efficient string matching: an aid to bibliographic search," *Communications of the ACM,* vol. 18, no. 6, pp. 333–340, 1975.
7. R. S. Boyer and J. S. Moore, "A fast string searching algorithm," *Communications of the ACM,* vol. 20, no. 10, pp. 762–772, 1977.
8. M. Ciubotariu, "Virus Cryptoanalysis," *Virus Bulletin,* 2003.
9. S. Potluri and C. Diedrich, "High Performance Intrusion Detection and Prevention Systems: A Survey," in *ECCWS2016-Proceedings fo the 15th European Conference on Cyber Warfare and Security*, 2016.
10. G. Vasiliadis, S. Antonatos, M. Polychronakis, E. P. Markatos and S. Ioannidis, "Gnort: High performance network intrusion detection using graphics processors," in *Recent Advances in Intrusion Detection*, Berlin/Heidelberg, Springer, 2008, pp. 116–134.
11. C.-H. Lin, C.-H. Liu, L.-S. Chien and a. S.-C. Chang, "Accelerating pattern matching using a novel parallel algorithm on GPUs," *IEEE Transactions on Computers,* vol. 62, no. 10, pp. 1906–1916, 2013.
12. Tumeo, O. Villa and D. Sciuto, "Efficient pattern matching on GPUs for intrusion detection systems," in *Proceedings of the 7th ACM international conference on Computing frontiers*, 2010.
13. Pungila and V. Negru, "A highly-efficient memory compression approach for GPU-accelerated virus signature matching," in *International Conference on Information Security*, 2012.
14. D. R. V. L. B. Thambawita, R. Ragel and D. Elkaduwe, "To use or not to use: Graphics processing units (GPUs) for pattern matching algorithms," in *Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on*, Colombo, Sri Lanka, 2014.
15. G. Vasiliadis and S. Ioannidis, "Gravity: a massively parallel antivirus engine," *International Workshop on Recent Advances in Intrusion Detection,* vol. 63, no. 7, pp. 79–96, 2010.
16. S. Ashkiani, N. Amenta and J. D. Owens, "Parallel Approaches to the String Matching Problem on the GPU," *Proceedings of the 28th ACM Symposium on Parallelism in Algorithms and Architectures,* pp. 275–285, 2016.
17. H. A. Kadhim and N. A. Rashid, "Parallel GPU-Based Hybrid String Matching Algorithm," *Advanced Computer and Communication Engineering Technology,* pp. 1199–1208, 2016.
18. Y.-S. Lin, C.-L. Lee and Y.-C. Chen, "A Capability-Based Hybrid CPU/GPU Pattern Matching Algorithm for Deep Packet Inspection," *International Journal of Computer and Communication Engineering,* vol. 5, no. 5, pp. 321–330, 2016.
19. P. Rogers, "Heterogeneous system architecture overview," in *Hot Chips 25 Symposium (HCS), 2013 IEEE*, Stanford, CA, USA, 2013.
20. R. N. Horspool, "Practical fast searching in strings," *Software: Practice and Experience,* vol. 10, no. 6, pp. 501–506, 1980.
21. D. Lemire and O. Kaser, "Faster 64-bit universal hashing using carry-less multiplications," *Journal of Cryptographic Engineering,* vol. 6, no. 3, pp. 171–185, 2016.

22. R. Poston, “How large is a piece of Malware?,” 27 July 2010. [Online]. Available: https://nakedsecurity.sophos.com/2010/07/27/large-piece-malware/. [Accessed 20 December 2017].
23. T. Tian and C.-P. Shih, *Software techniques for shared-cache multi-core systems,* Intel Software Network, 2007.

Complete Design of a Hardware and Software Framework for PWM/Discrete PID-Based Speed Control of a Permanent-Magnet DC Motor Without Prior Knowledge of the Motor's Parameters

Chady El Moucary, Abdallah Kassem, Walid Zakhem, Chaybane Ghabach, Roger El Khoury, and Patrick Rizk

1 Introduction

The objective of this chapter resides in presenting a comprehensive approach for the design and implementation of an effective computational tool to control the speed of a Permanent-Magnet DC (PMDC) motor.

This type of motor is widely used in innumerable industrial applications due to its rugged structure, low cost, high efficiency, and pertinent characteristics. It does not require a separate excitation coil, hence the reduced size and lower power consumption. The general block diagram that depicts the overall system is shown in Fig. 1.

The speed control is achieved using a discrete PID controller with the ability of online tuning and adjusting the parameters to a changing desired trajectory. PID controllers are also widely used for their versatile features in monitoring the shape of the tracking error according to desired mode of responses, steady-state errors, and required dynamics. They are also chosen for their ease and variety of implementation techniques and methods.

The computational tool encompasses a PIC microcontroller and a fully comprehensive, user-friendly, and resourceful interface designed using Visual Basic. VB programming offers an adaptable platform known for its appealing interfacing

C. E. Moucary (✉) · A. Kassem · W. Zakhem · C. Ghabach · R. E. Khoury · P. Rizk
Department of Electrical, Computer, and Communication Engineering, Notre Dame University–Louaize, Zouk Mosbeh, Lebanon
e-mail: celmoucary@ndu.edu.lb; akassem@ndu.edu.lb; wzakhem@ndu.edu.lb

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_10

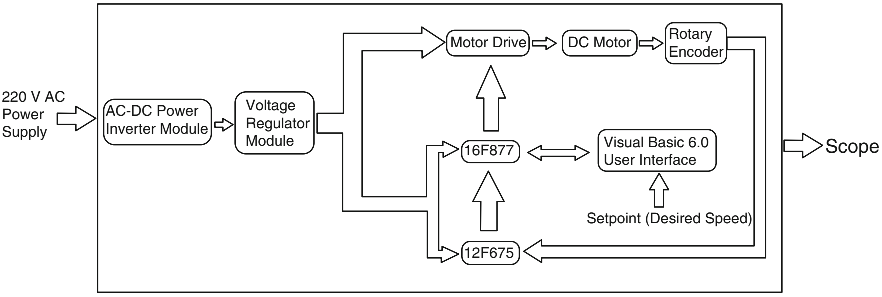


Fig. 1 Block diagram of the drive

features and portability. The GUI is fashionably designed to incorporate a broad and multipurpose portal to access, monitor, edit, and record all pertinent parameters of the entire drive.

This chapter will be divided in nine sections. Fundamentals of the speed control of a PDMC motor using PWM techniques are presented in the second section. The third section discusses the general model of the PID controller and the pertinent direct canonical forms are then presented in Sect. 4. The implementation of the PID controller is showcased in Sect. 5. Electric circuit design and pertinent schematics are then elaborated in Sect. 6, which also shows the platform used for simulation and testing purposes. The piloting software and the functionality of the GUI are overviewed in Sect. 7, followed by experimental results depicted in Sect. 8 that underline the effectiveness of the overall framework. Finally, a conclusive summary is presented in Sect. 9.

2 Pulse Width Modulation (PWM)

The foundation of modern mechanical systems lies in control systems that allow designing of apparatus that would theoretically perform according to any granularity in terms of specification requirements such as dynamic and steady-state behaviors. Control operations can be achieved in either open-loop or closed-loop approaches; the key difference is feedback. In open-loop configurations, the system acts completely on the basis of input; the output has no effect on the underlying action. Hence, no feedback is available and/or used to adjust the behavior to the desired path or outcome. Arguably the most ingenious tool in this case is the closed-loop outlook, which shares the most constituent components of the aforementioned open-loop platform but with some relevant data being passed back from some point into the control system to another preceding point. Such data is primarily used to modify/correct the response for the actual output to closely follow a desired reference trajectory in terms of many related criteria; hence, the system becomes self-adjusting. Despite the fact that open-loop systems are by far simpler to design

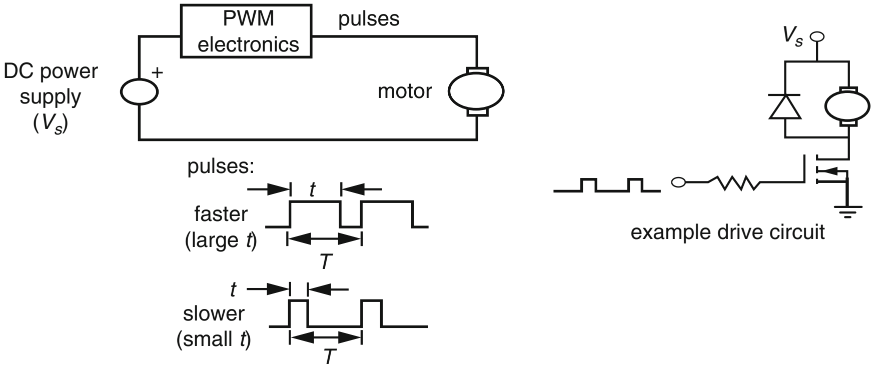


Fig. 2 Pulse Width Modulation for a DC motor

and entail low implementation cost, closed-loop systems offer a decisively more resourceful framework, and somehow become mandatory when strict requirements in terms of dynamic and steady-state operations are needed. Closed-loop speed control systems involve speed sensors such as rotary encoders that make the actual output available for feedback. Actual speed measurement values are constantly compared to the desired reference value, which is set by the user and called the *set point*. The difference between those values is then fed to an astutely designed controller that would adjust the motor input (voltage or current) to ensure effective tracking of the set point. Electronic speed controllers are of two types: linear amplifiers and pulse width modulators (PWMs). PWM controllers present the advantage of either driving bipolar power transistors rapidly between cutoff and saturation or turning FETs ON and OFF. In either case, power dissipation is small. Servo amplifiers using linear power amplification are satisfactory but produce a lot of heat, because they function in the transistor linear region. Commercial servo controllers can be achieved using linear amplifiers, but because of lower power requirements, ease of design, smaller size, and lower cost, switched amplifier designs are used [1–4].

Figure 2 depicts the principle of a PWM amplifier. A DC power supply voltage is rapidly switched at a fixed frequency f between two values (e.g., ON and OFF). This frequency is often in excess of 1 kHz. The high value is held during a variable pulse width t within the fixed period T where $T = 1/f$.

The resulting asymmetric waveform has a duty cycle defined as the ratio between the ON time and the period of the waveform, usually specified as a percentage:

$$\text{Duty cycle} = \frac{t}{T} \times 100 \tag{1}$$

As the duty cycle is changed (by the controller), the average current through the motor changes, causing changes in speed and torque at the output. It is primarily the duty cycle, and not the value of the power supply voltage, that is used to control the

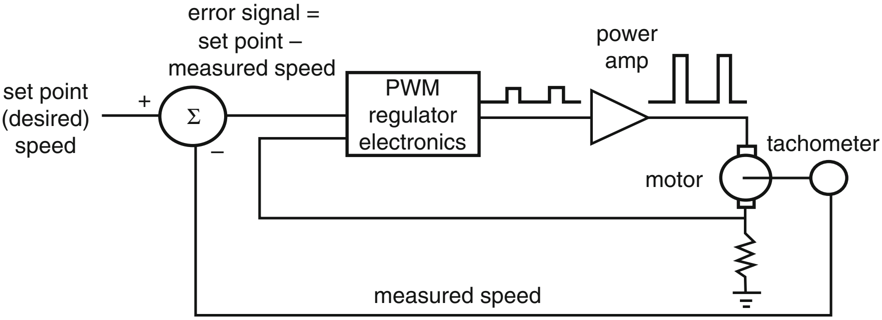


Fig. 3 PWM speed control

speed of the motor. The block diagram of a PWM speed feedback control system for a DC motor is shown in Fig. 3. A voltage tachometer produces an output linearly related to the motor speed. This is compared to the desired speed set point (another voltage that can be manually set or computer controlled). The error and the motor current are sensed by a pulse-width-modulation regulator that produces a width-modulated square wave as an output. This signal is amplified to a level appropriate to drive the motor.

In a PWM motor controller, the armature voltage switches rapidly, and the current through the motor is affected by the motor inductance and resistance. Since the switching speed is high, the resulting current through the motor has a small fluctuation around an average value. As the duty cycle grows larger, the average current grows larger and the motor speed increases.

3 PID Controller

The proportional integral derivative or the PID is one of the most commonly used controllers nowadays. The control signal is generated from three terms: a term that is proportional to the error, a term that is integral to the error, and a term that is derivative of the error [5–8]. The sum of these three terms will serve as control signal for the PWM block. The PID controller can be modeled by the block diagram shown in Fig. 4.

The components of a PID system are:

- Proportional term K_P depends on the present error. This term defines the speed of change in the output.
- Integral term K_i is the accumulation of past errors. It aids in reaching with a faster response the steady state and eliminates the residual steady-state error that results from the pure use of a proportional controller.
- Derivative term K_d is the prediction of future errors. This term of control is used to decrease the magnitude of the overshoot.

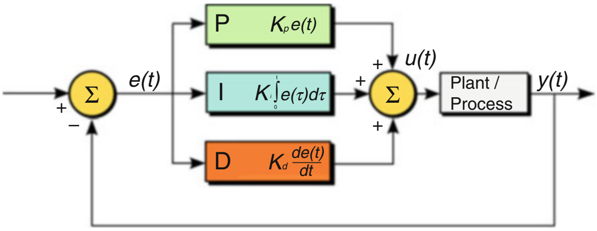


Fig. 4 PID block diagram

Analyzing the block diagram, we obtain the following equation:

$$u(t) = K_p e(t) + K_i \int_0^t e(t)\mathrm{d}t + K_d \frac{\mathrm{d}e(t)}{\mathrm{d}t} \tag{2}$$

where $K_i = \frac{K_p}{T_i}$ and $K_d = K_p T_d$; therefore, we can write the transfer function of a continuous-time PID as

$$\frac{U(s)}{E(s)} = K_p + \frac{K_p}{T_i s} + K_p T_d s \tag{3}$$

The discrete form of PID controller can also be derived by finding the z-transform of equation (3).

Therefore, we obtain:

$$\frac{U(z)}{E(z)} = K_p \left[1 + \frac{T}{T_i \left(1 - z^{-1}\right)} + T_d \frac{\left(1 - z^{-1}\right)}{T} \right] \tag{4}$$

Then

$$\begin{aligned} u(kT) &= u\,(KT - T) + K_p\,[e(kT) - e\,(kT - T)] \\ &+ \frac{K_p T_d}{T}\,[e(kT) - 2e\,(kT - T) - e\,(kT - 2T)] \end{aligned} \tag{5}$$

The PID controller is accurately tuned using the Ziegler–Nichols approach [9–13]. The closed-loop tuning algorithm based on plant closed-loop tests is as follows:

1. *Disable any derivative and integral action in the controller and leave only the proportional action.*
2. *Carry out a set-point step test and observe the system response.*
3. *Repeat the set-point test with increased (or decreased) controller gain until a stable oscillation is achieved; this gain is called the ultimate gain* $\boldsymbol{K_u}$.
4. *Read the period of the steady oscillation and let this be* $\boldsymbol{T_u}$.

Table 1 Effect of PID parameters on the tracking error

Parameter	Rise time	Overshoot	Settling time	Steady-state error	Stability
K_p	Decrease	Increase	Small change	Decrease	Degrade
K_i	Decrease	Increase	Increase	Eliminate	Degrade
K_d	Minor change	Decrease	Decrease	No effect in theory	Improve if small

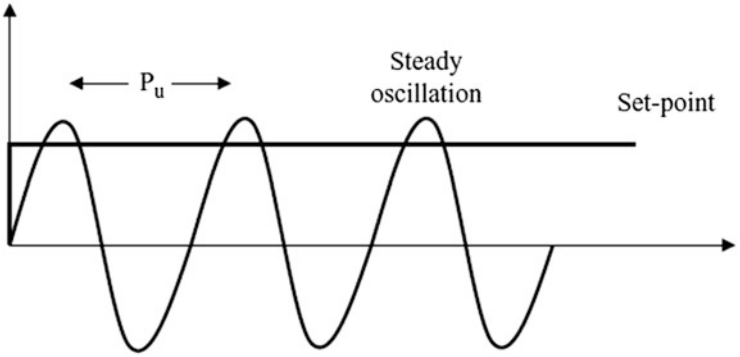


Fig. 5 Ziegler–Nichols closed loop

5. *Calculate the controller parameters according to the following formulas*: $\boldsymbol{K_P = 0.45K_u}$, $\boldsymbol{T_i = T_u/1.2}$ *in case of PI controller, and* $\boldsymbol{K_P = 0.6K_u}$, $\boldsymbol{T_i = T_u/2}$, $\boldsymbol{T_d = T_u/8}$ *in the case of the PID controller.*

Table 1 summarizes the effect of changing the parameters of the PID controller (Fig. 5).

The Ziegler–Nichols rules are then applied to obtain initial controller design followed by design iteration and refinement. By using the Ziegler–Nichols formulas we have:

$K_P = 0.6, K_U = 53.13, K_I = 1.2, K_U/T_U = 1280.2$, and $K_D = 0.6, K_U\ T_U/8 = 55$. Those parameters yielded a settling time of about 2.4 s and a percentage overshoot of approximately 60%.

The Ziegler method based on assumed forms of the process presents decisive advantages for it allows achieving a speed control of the motor without prior knowledge of its parameters. This method shall procure the system with the required constants $a0$, $a1$, $a2$, $b1$, $b2$ that will be used as preset values for the compensator. The physical meaning of these constants and their relationship to the PID parameters are explained in the following section.

4 Direct Canonical Forms

The strategy of a numerical control structure begins with a precise model of the process to be controlled. Then a control algorithm is developed, which will ensure the required system response. The loop is closed by using a digital computer as the controller. The computer implements the control procedure in order to obtain the desired response. Different approaches do have dissimilar computational efficiencies, dissimilar sensitivities to parameter errors, and dissimilar programming techniques are needed in each case. As such, various approaches are available for implementation. To name a few, cascaded structures, parallel structures, second-order structures, and direct structures could be used for building the controller. In fact, two different types of direct structures can be considered: the direct noncanonical structure and the direct canonical structure [14–17].

The direct canonical structure is selected since it presents significant advantages over other structures such as memory size and efficiency as it requires a smaller number of memory locations and the number of delay elements is fixed.

In direct structure, the coefficients a_j and b_j appear as multipliers. By considering that b_0=1, for the discrete compensator, therefore, we can express

$$D(z) = \frac{U(z)}{E(z)} = \frac{\sum_{j=0}^{n} a_j z^{-j}}{1 + \sum_{j=0}^{n} b_j z^{-j}} \tag{6}$$

Let us introduce now a new variable R(z) such that

$$\frac{U(z)}{R(z)} \frac{R(z)}{E(z)} = \frac{\sum_{j=0}^{n} a_j z^{-j}}{\sum_{j=0}^{n} b_j z^{-j}} \tag{7}$$

or

$$\frac{U(z)}{R(z)} = \sum_{j=0}^{n} a_j z^{-j} \text{ and } \frac{E(z)}{R(z)} = \sum_{j=0}^{n} b_j z^{-j} \tag{8}$$

Assume that the transfer function of a digital controller is

$$R(z) = E(z) - \sum_{j=1}^{n} b_j z^{-j} R(z) \tag{9}$$

$$\text{And } U(z) = \sum_{j=0}^{n} a_j z^{-j} R(z) \tag{10}$$

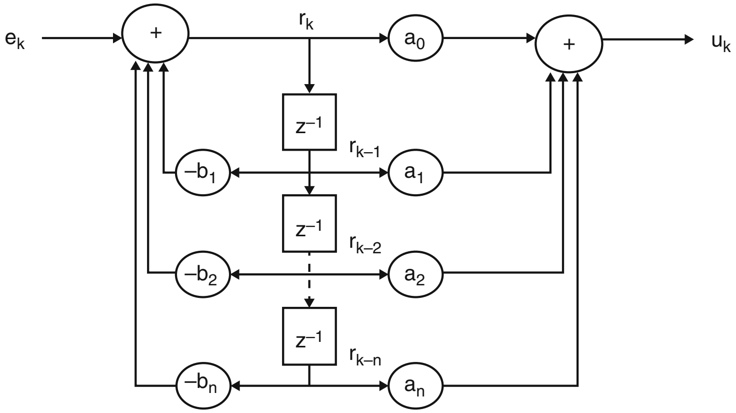


Fig. 6 Block diagram implementation discrete compensator

The equations above can be written in time domain

$$r_k = e_k - \sum_{j=1}^{n} b_j r_{k-j} \tag{11}$$

$$u_k = \sum_{j=0}^{n} a_j r_{k-j} \tag{12}$$

Those equations define the direct form, and the block diagram of implementation is shown in Fig. 6. The controller is made up of delays, adders, and multipliers.

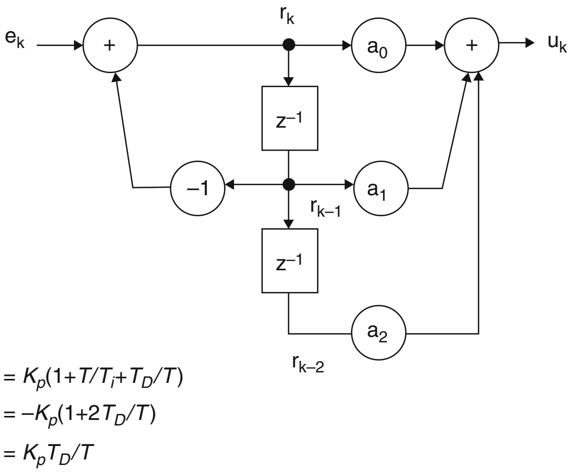
5 Controller Implementation

The z-transform of the PID controller was derived before, and is reproduced here for convenience:

$$D(z) = K_p + \frac{K_p T}{T_i \left(1 - z^{-1}\right)} + \frac{K_p T_d \left(1 - z^{-1}\right)}{T} \tag{13}$$

An alternative implementation of the PID would be to find a second-order transfer function for $D(z)$ and then use the direct structure to implement it. The equation can be written as

$$D(z) = \frac{K_p \left(1 - z^{-1}\right) + \frac{K_p T}{T_i} + \left(\frac{K_p T_d}{T}\right) \left(1 - z^{-1}\right)^2}{1 - z^{-1}} \tag{14}$$



$\alpha_0 = K_p(1{+}T/T_i{+}T_D/T)$

$\alpha_1 = -K_p(1{+}2T_D/T)$

$\alpha_2 = K_pT_D/T$

Fig. 7 PID implementation—direct canonical structure

Therefore,

$$D(z) = \frac{K_p + \frac{K_pT}{T_i} + \frac{K_pT_d}{T} - \left(K_p + \frac{2K_pT_d}{T}\right)z^{-1} + z^{-2}\left(K_pT_d/T\right)}{1 - z^{-1}}, \tag{15}$$

which is of the form $\frac{a_0+a_1z^{-1}+a_2z^{-2}}{1+b_1z^{-1}+b_2z^{-2}}$,where

$$a_0 = K_p\left(1 + \frac{T}{T_i} + \frac{T_d}{T}\right), a_1 = -K_p\left(1 + \frac{2T_d}{T}\right), a_2 = \frac{K_pT_d}{T}, b_1 = -1, b_2 = 0. \tag{16}$$

Figure 7 shows the PID implementation as direct canonical structure.

Considering again the velocity form of PID, and replacing the *k*T simply by subscript *k*, we can write

$$u_k = u_{k-1} + \left[K_p + \frac{K_pT}{T_i} + \frac{K_pT_d}{T}\right]e_k - \left[K_p + \frac{2K_pT_d}{T}\right]e_{k-1} + \frac{K_PT_d}{T}e_{k-2} \tag{17}$$

Alternatively, we can write (17) in a simpler form as

$$u_k = u_{k-1} + ae_k + be_{k-1} + ce_{k-2} \tag{18}$$

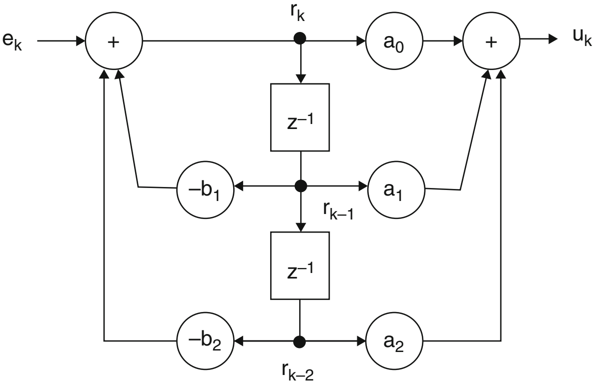


Fig. 8 Second-order module implementations

where

$$a = K_p + \frac{K_p T}{T_i} + \frac{K_p T_d}{T}, b = -\left[K_p + \frac{2K_p T_d}{T}\right], c = \frac{K_p T_d}{T} \tag{19}$$

By taking the z-transform we obtain

$$D(z) = \frac{U(z)}{E(z)} = \frac{a + bz^{-1} + cz^{-2}}{1 - z^{-1}} \tag{20}$$

Notice that if only proportional plus integral (PI) action is required, the derivative constant T_d can be set to zero and the PI equation becomes $D(z) = \frac{U(z)}{E(z)} = \frac{a_1+bz^{-1}}{1-z^{-1}}$with $a = K_p + \frac{K_pT}{T_i}$ and $b = K_p$.

The second-order module is shown in Fig. 8. This module serves as our final model of the discrete controller. It ensures an enhanced response in presence of disturbances.

Where

$$Q(z) = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \tag{21}$$

The difference equations describing such a module are

$$r_k = e_k - b_1 r_{k-1} - b_2 r_{k-2} \tag{22}$$

$$u_k = a_0 r_k + a_1 r_{k-1} + a_2 r_{k-2} \tag{23}$$

If we let

$$M_1 = -b_1 r_{k-1} - b_2 r_{k-2} \text{and } M_2 = a_1 r_{k-1} + a_2 r_{k-2} \quad (24)$$

Then these equations for the second-order module become

$$r_k = e_k + M_1 \quad (25)$$

$$u_k = a_0 r_k + M_2 \quad (26)$$

6 Circuit Design and Schematics

In this section, we shall discuss the electric circuit implementation of the system using a bottom-up approach. The function of each component will be analyzed, then the overall operation is summarized. Figure 9 depicts the complete schematics of the drive.

In order to minimize the number of wires needed, we implemented a voltage regulator module, shown in Fig. 10, which controls various components. Along with the 30 V power supply, a 5 V DC source is needed in order to supply the PIC16F877A as well as PIC12F675. Moreover, a 12 V voltage source is required to supply the power MOSFET transistor.

The PIC16F877A has 40 pins and five ports. It uses reduced instruction set RISC (35-instruction set); it can support four different types of oscillators—the crystal

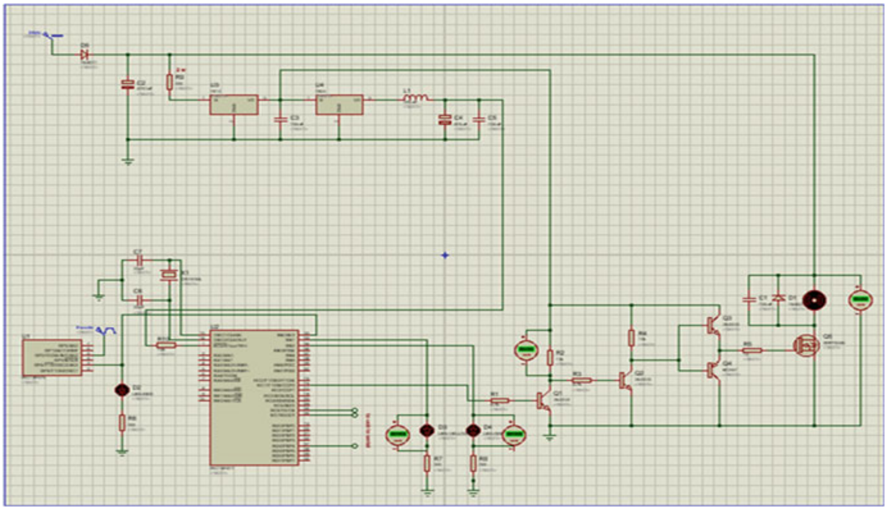


Fig. 9 Circuit of DC motor control using PID

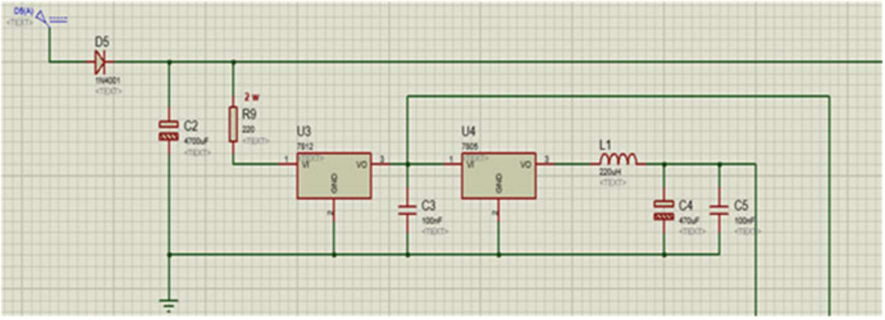


Fig. 10 Voltage regulator module

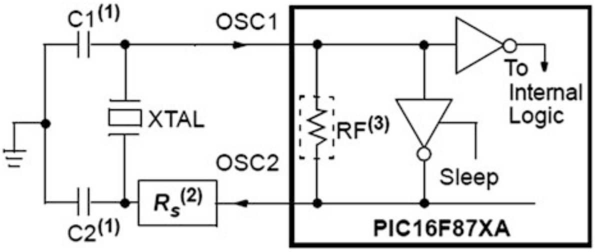


Fig. 11 Crystal oscillator

oscillator 770(XT) was adopted. Figure 11 shows the choice of the capacitors of 13 pF each in order to generate a frequency of 4 MHz, and since most instructions require four cycles, the operating frequency of the PIC is 1 MHz. Moreover, it has three different timers and can accommodate up to 15 different sources of interrupts. Only three interrupts are required: INTERRUPT_TIMER0, INTERRUPT_RC, and INTERRUPT_RB0. These three interrupt service routines are assigned to Timer0, USART (universal asynchronous transmitter receiver), and pin RB0, respectively.

INTERRUPT_TIMER0 subroutine is set each 100 ms, the flag TO1F will be set and thus, the interrupt service routine is called. This routine will read the actual speed of the motor and will place it in a variable called ACTUAL_SPEED.

To measure the speed of the motor, a rotary encoder was used; each round per second (or rps) will generate 400 pulses. The rotary encoder is connected to the input of the PIC12F675 (GP2) as per Fig. 12. By doing so, all pulses are saved and no data is lost. The microcontroller divides the number of pulses obtained from the rotary encoder by 40, that is, the number of pulses is now 400/40. Thus, each one round from the motor's shaft is now equivalent to 10 pulses. This relationship summarizes the measuring procedure of the speed as a linear relationship for the calculation of the equivalent number of pulses for *x* number of physical pulses.

To make sure that we tackled the issue of speed measuring (all pulses are counted) let us consider a high-speed scenario. The rated speed of the motor is 3300 rpm, which will generate (3300 × 400)/60 or 22,000 pulses/s. This number

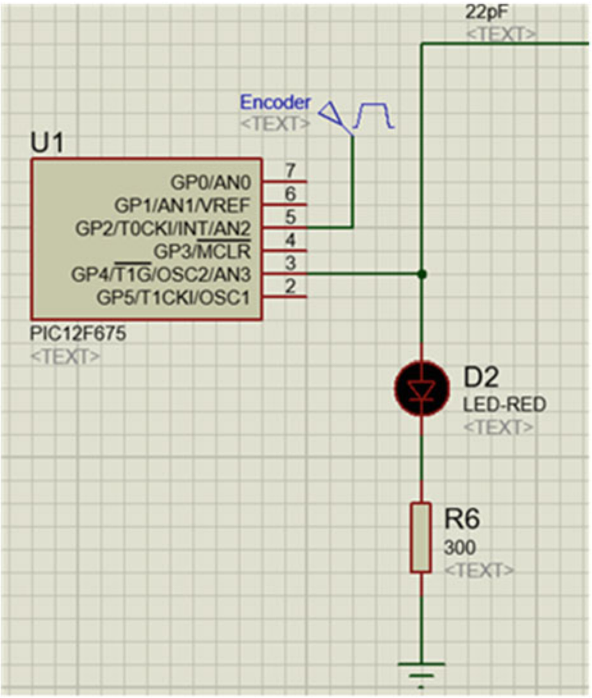


Fig. 12 PIC12F675 module

will be divided by 40 and thus, if the motor is running at its maximum speed (3300 rpm or 55 rps) the total number of pulses generated will be 550 pulses per second. However, the PIC16F877A is an 8-bit microcontroller and hence, could not accommodate the entire range of numbers issued by the pulse counter. Consequently, an interrupt timer0 is set at each 100 ms and hence, instead of reading 550 pulses per second at rated speed, only 55 pulses per 100 ms will be read. Because this is the maximum number of pulses that can be generated; therefore, if the motor is running at any other given speed, the number of pulses will all be saved without any loss since an 8-bit register can save up to $2^8 - 1$ or 255.

Figure 13 shows a Totem Pole driver circuit configuration chosen over the H-Bridge driver due to its high abilities for fast switching as well as for having a low cost and being easily implementable.

Figure 14 is a standard rectifier circuit that is simulated using National Instruments Multisim.

Figure 15 shows the 3D models of the PCB circuit generated using Proteus.

Figure 16 shows the actual image of the PCB with the components installed in it.

All of the above-detailed circuits were tested throughout simulation to check the readiness of the framework and its compliance to the system's specification requirements before any experiment is carried out. This step is crucial in order to

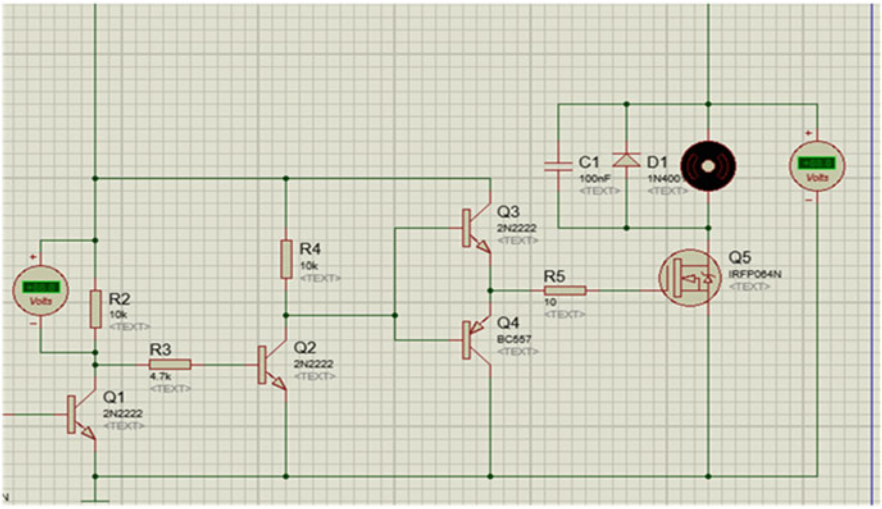


Fig. 13 The drive motor module

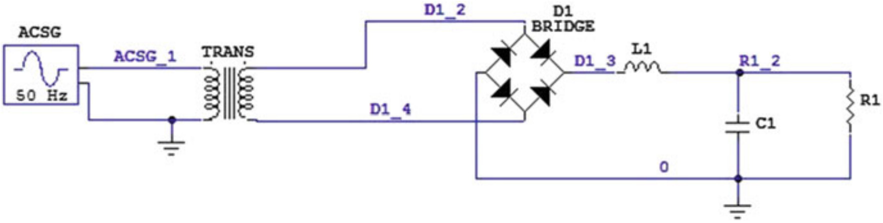


Fig. 14 Power supply module

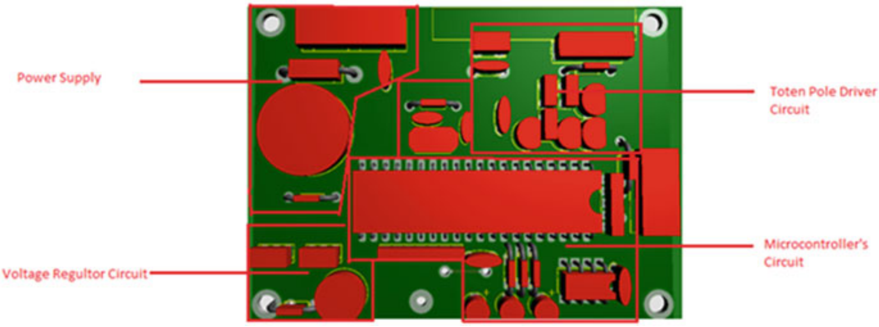


Fig. 15 PCB layout

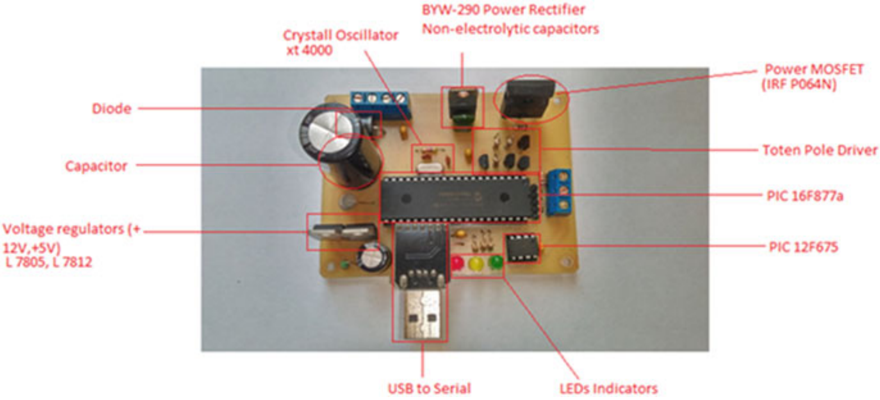


Fig. 16 PCB with components installed (top view)

prevent any possible damage to the motor and/or the electric components. The next step is to develop the pilot software to be used for the control scheme. A resourceful GUI is presented in the following section that offers a versatile and effective tool, as will be shown.

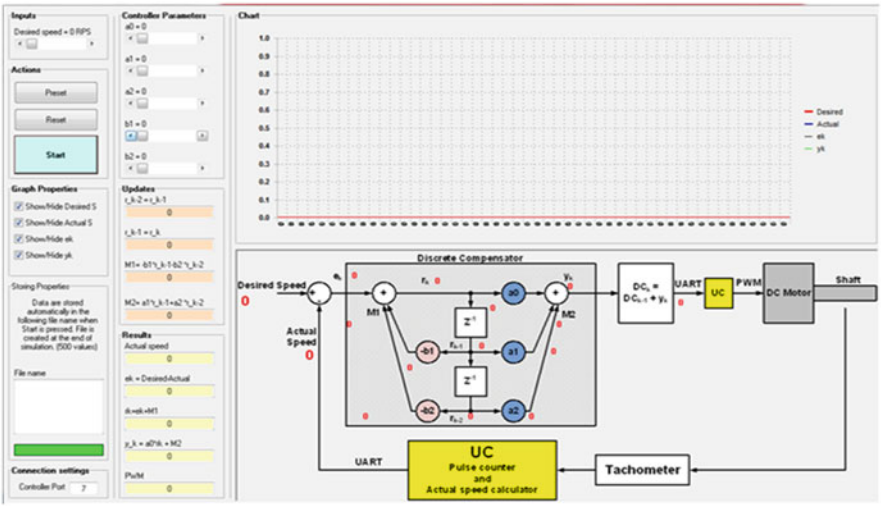
7 Software Design

In this section, the main pilot software is presented. It is developed in Visual Basic 6, which is an integrated development environment (IDE) that was developed by Microsoft. A Graphical User Interface was developed to entail the user a friendly yet versatile and effective tool for parameter calibration and control. The GUI is depicted in Fig. 17. It is divided into several parts: the controller port interface with the main computer, the compensator parameters, and the display properties, as showcased in Figs. 18 and 19.

To obtain the most optimum response in terms of overshoot as well as settling time, a “Preset” button is available to set the values of $a0$, $a1$, $a2$, $b1$, $b2$ to 1.3, 0.5, 0.2, 0.1, and 0.6, respectively. In fact, these values were obtained after extensive PID tuning.

Figure 20 below showcased a closed look at the compensator structure.

Figure 21 depicts where the online pertinent parameters are calculated and displayed. It also exhibits the capacity of saving results and data in excel files with all relevant parameters for offline analysis.



Inputs

Desired speed = 0 RPS

Controller Parameters

a0 = 0

a1 = 0

a2 = 0

b1 = 0

b2 = 0

Graph Properties

☑ Show/Hide Desired S

☑ Show/Hide Actual S

☑ Show/Hide ek

☑ Show/Hide yk

Storing Properties

Data are stored automatically in the following file name when Start is pressed. File is created at the end of simulation. (500 values)

File name

Connection settings

Controller Port 7

Updates

r_k-2 = r_k-1 0

r_k-1 = r_k 0

M1= -b1*r_k-1-b2*r_k-2 0

M2= a1*r_k-1+a2*r_k-2 0

Results

Actual speed 0

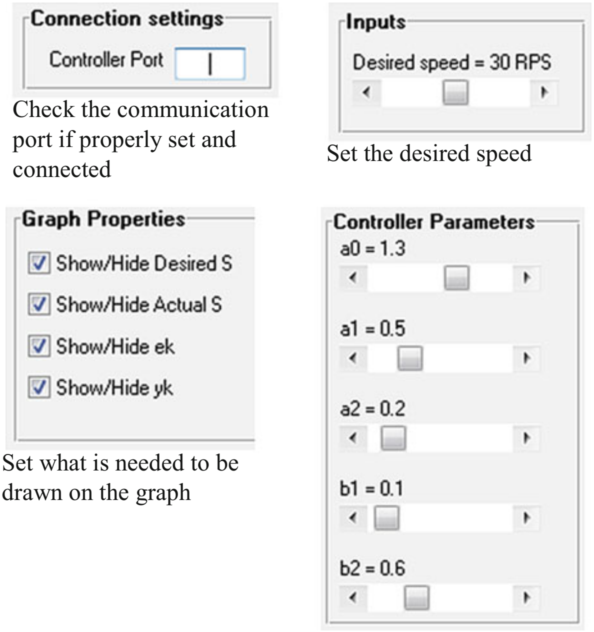
ek = Desired-Actual 0

rk=ek+M1 0

y_k = a0*rk + M2 0

PWM 0

Fig. 17 Comprehensive GUI interface



Connection settings

Controller Port |

Check the communication port if properly set and connected

Inputs

Desired speed = 30 RPS

Set the desired speed

Graph Properties

☑ Show/Hide Desired S

☑ Show/Hide Actual S

☑ Show/Hide ek

☑ Show/Hide yk

Set what is needed to be drawn on the graph

Controller Parameters

a0 = 1.3

a1 = 0.5

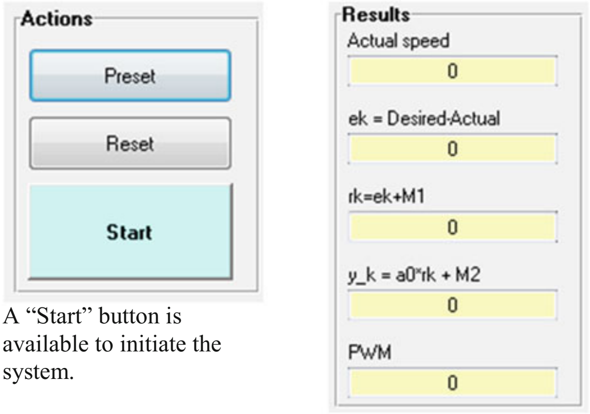
a2 = 0.2

b1 = 0.1

b2 = 0.6

Set the controller parameters.

Fig. 18 Display properties and compensator parameters



Results

Actual speed
0

ek = Desired-Actual
0

rk=ek+M1
0

y_k = a0*rk + M2
0

PWM
0

A “Start” button is available to initiate the system.

The calculated results will be shown

Fig. 19 Compensator options and tuning

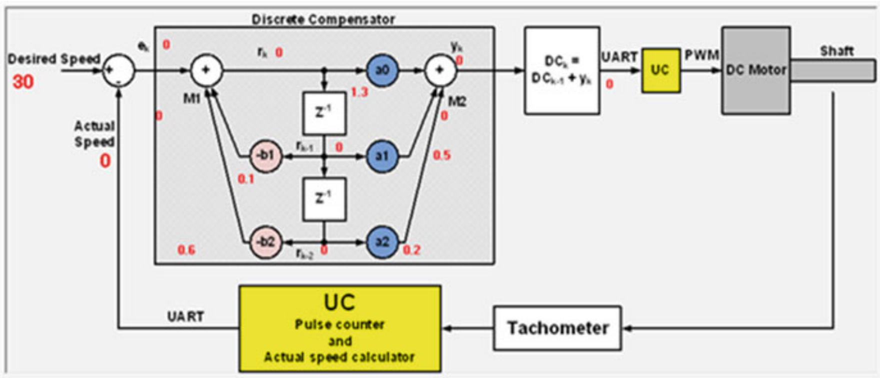
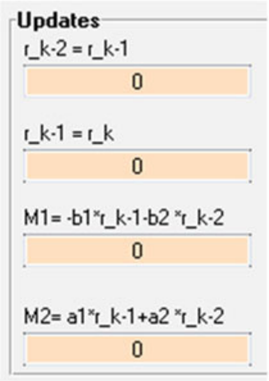


Fig. 20 Discrete compensator

8 Experimental Results

Several experimental tests were carried out that demonstrated the effectiveness of the overall framework in terms of the dynamic and steady-state behavior of the system. Typical situations are presented below. The first one involves a speed control with no load disturbances. Figure 22 shows the motor’s response to a step in speed set at 30 rps. The behavior is quite satisfactory and overshoot is noted as previewed by the compensator’s parameters.

The second typical situation involves the application of a sudden load or disturbance at different instants. Figure 23 clearly shows that the compensator was able to handle the speed control by keeping track of the desired speed after slight



Updates

r_k-2 = r_k-1

0

r_k-1 = r_k

0

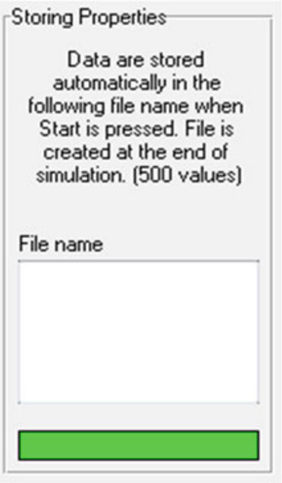
M1= -b1*r_k-1-b2 *r_k-2

0

M2= a1*r_k-1+a2 *r_k-2

0

Internal calculation of the system in order to generate the corresponding PWM.



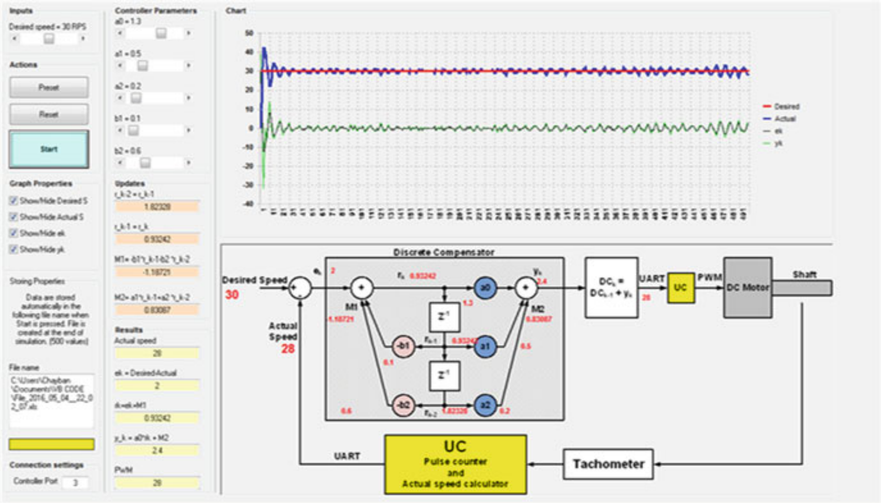
Storing Properties

Data are stored automatically in the following file name when Start is pressed. File is created at the end of simulation. (500 values)

File name

The results of the experiment are saved in a excel file in which the user can check after the simulation finishes (500 samples)

Fig. 21 Online computation of related parameters; saving results



Inputs

Desired speed = 30 RPS

Actions

Preset

Reset

Start

Graph Properties

☑ Show/Hide Desired S

☑ Show/Hide Actual S

☑ Show/Hide ek

☑ Show/Hide yk

Storing Properties

Data are stored automatically in the following file name when Start is pressed. File is created at the end of simulation. (500 values)

File name

C:\Users\Chayban \Documents\VB CODE \File_2016_05_04__22_0 2_07.xls

Connection settings

Controller Port 3

Controller Parameters

a0 = 1.3

a1 = 0.5

a2 = 0.2

b1 = 0.1

b2 = 0.6

Updates

r_k-2 = r_k-1

1.82328

r_k-1 = r_k

0.93242

M1= -b1*r_k-1-b2 *r_k-2

-1.18721

M2= a1*r_k-1+a2 *r_k-2

0.83067

Results

Actual speed

28

ek = Desired-Actual

2

rk=ek+M1

0.93242

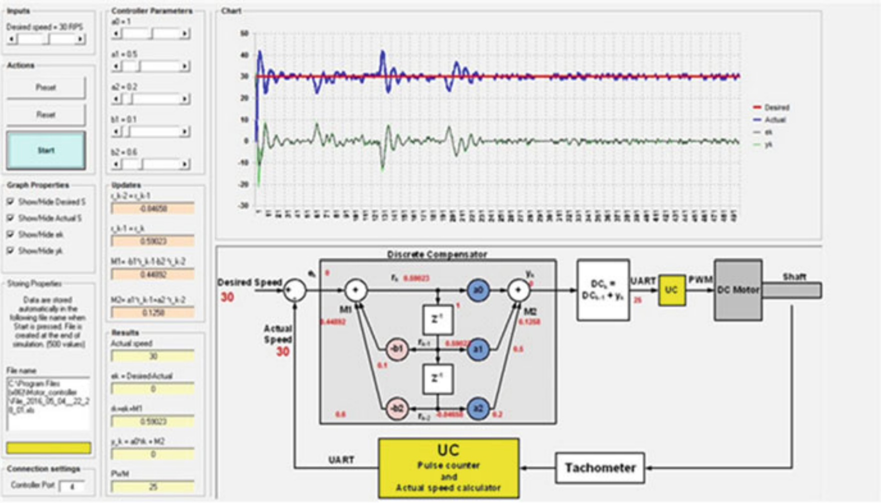
y_k = a0*rk + M2

2.4

PWM

28

Fig. 22 Speed control (no load applied)



Inputs
Desired speed = 30 RPS

Controller Parameters
a0 = 1
a1 = 0.5
a2 = 0.2
b1 = 0.1
b2 = 0.6

Graph Properties
☑ Show/Hide Desired S
☑ Show/Hide Actual S
☑ Show/Hide ek
☑ Show/Hide yk

Storing Properties
Data are stored automatically in the following file name when Start is pressed. File is created at the end of simulation. (500 values)

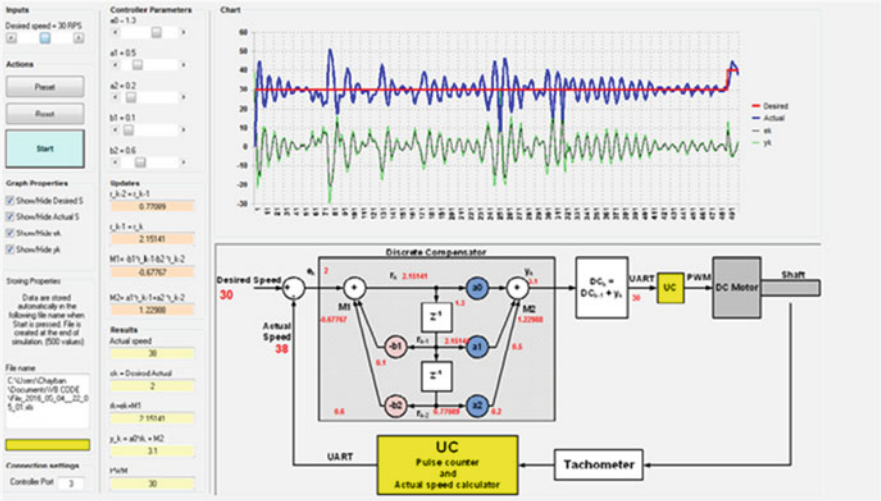
File name
C:\Program Files (x86)\Motor_controller\File_2016_05_04__22_20_01.xls

Connection settings
Controller Port 4

Updates
r_k-2 = r_k-1: -0.04658
r_k-1 = r_k: 0.59023
M1= -b1*r_k-1-b2*r_k-2: 0.44092
M2= a1*r_k-1+a2*r_k-2: 0.1258

Results
Actual speed: 30
ek = Desired-Actual: 0
rk=ek+M1: 0.59023
y_k = a0*rk + M2: 0
PWM: 25

Fig. 23 Speed control with load disturbance



Inputs
Desired speed = 30 RPS

Controller Parameters
a0 = 1.3
a1 = 0.5
a2 = 0.2
b1 = 0.1
b2 = 0.6

Graph Properties
☑ Show/Hide Desired S
☑ Show/Hide Actual S
☑ Show/Hide ek
☑ Show/Hide yk

Storing Properties
Data are stored automatically in the following file name when Start is pressed. File is created at the end of simulation. (500 values)

File name
C:\Users\Chayban\Documents\VB CODE\File_2016_05_04__22_05_01.xls

Connection settings
Controller Port 3

Updates
r_k-2 = r_k-1: 0.77089
r_k-1 = r_k: 2.15141
M1= -b1*r_k-1-b2*r_k-2: -0.67767
M2= a1*r_k-1+a2*r_k-2: 1.22988

Results
Actual speed: 38
ek = Desired-Actual: 2
rk=ek+M1: 2.15141
y_k = a0*rk + M2: 3.1
PWM: 30

Fig. 24 Speed control (heavy load)

deviation and oscillations. In fact, new PWM values were generated in order to adjust the speed; those values are shown online via the GUI and then saved with the related speed curves.

Figure 24 shows a heavier load was applied to the shafts of the motor. The motor would still run at the desired speed. However, since the applied load was heavy (the shafts were almost about to stop as per the above graph), greater oscillations are observed but the motor did not halt and the average speed was maintained.

9 Conclusion

In this chapter, we developed a comprehensive design of a hardware and software framework for PWM/Discrete PID-based speed control of a Permanent-Magnet DC Motor without prior knowledge of the motor's parameters. Ziegler–Nichols approach associated with Direct Canonical Forms theory allowed tuning a discrete PID compensator to achieve speed control that is robust to load torque disturbances and able to meet requirements in terms of steady-state error, time response, and percent overshoot. Simulation and experimental results showcased the effectiveness of the design that is low cost and somehow simple to implement.

References

1. Buja, G.S. and Kazmierkowski, M.P., 2004. Direct torque control of PWM inverter-fed AC motors-a survey. IEEE Transactions on industrial electronics, 51(4), pp.744–757.
2. Ogasawara, S., Akagi, H. and Nabae, A., 1990. A novel PWM scheme of voltage source inverters based on space vector theory. Archiv für Elektrotechnik, 74(1), pp.33–41.
3. Kazmierkowski, M.P. and Malesani, L., 1998. Current control techniques for three-phase voltage-source PWM converters: A survey. IEEE Transactions on industrial electronics, 45(5), pp.691–703.
4. Öztürk, N., Kaplan, O. and Çelik, E., 2017. Zero-current switching technique for constant voltage constant frequency sinusoidal PWM inverter. Electrical Engineering, pp.1–11.
5. Leon, J.I., Kouro, S., Franquelo, L.G., Rodriguez, J. and Wu, B., 2016. The essential role and the continuous evolution of modulation techniques for voltage-source inverters in the past, present, and future power electronics. IEEE Transactions on Industrial Electronics, 63(5), pp.2688–2701.
6. Johnson, M.A. and Moradi, M.H., 2005. PID control. Springer-Verlag London Limited.
7. Hamamci, S.E., 2008. Stabilization using fractional-order PI and PID controllers. Nonlinear Dynamics, 51(1), pp.329–343.
8. Vilanova, R. and Visioli, A., 2012. PID control in the third millennium. London: Springer.
9. Lin, C.L. and Jan, H.Y., 2005. An approach to solving for multi-objective optimization problem with application to linear motor control design. Control and intelligent systems, 33(2), pp.75–86.
10. Hendy, H., Rui, X., Zhou, Q. and Khalil, M., 2014. Controller parameters tuning based on transfer matrix method for multibody systems. Advances in Mechanical Engineering, 6, p.957684.
11. Mudi, R.K. and Pal, N.R., 1999. A robust self-tuning scheme for PI-and PD-type fuzzy controllers. IEEE Transactions on fuzzy systems, 7(1), pp.2–16.
12. Åström, K.J. and Hägglund, T., 2004. Revisiting the Ziegler–Nichols step response method for PID control. Journal of process control, 14(6), pp.635–650.
13. Åström, K.J. and Hägglund, T., 1984. Automatic tuning of simple regulators with specifications on phase and amplitude margins. Automatica, 20(5), pp.645–651.
14. Barbosa, R.S., Machado, J.T. and Ferreira, I.M., 2004. Tuning of PID controllers based on Bode's ideal transfer function. Nonlinear dynamics, 38(1), pp.305–321.
15. Åström, K.J. and Hägglund, T., 2006. PID control. IEEE CONTROL SYSTEMS MAGAZINE, 1066(033X/06).

16. Chemuturi, Murali. Requirements Engineering and Management for Software Development Projects. New York: Springer, 2013.
17. Carriegos, M. and Hermida-Alonso, J.A., 2003. Canonical forms for single input linear systems. Systems & control letters, 49(2), pp.99–110.

Analysis of DDoS Attack-Aware Software-Defined Networking Controller Placement in Malaysia

Muhammad Reazul Haque, Saw Chin Tan, Ching Kwang Lee, Zulfadzli Yusoff, Sameer Ali, Ir. Rizaludin Kaspin, and Salvatore Renato Ziri

1 Introduction

The kernel brainchild of software-defined network (SDN) is to separate the control plane and the data plane [1] by creating special software that can be operated under separate hardware [2]. It is necessary to introduce SDN in modern telecommunication and computer networks as they became more sophisticated and more difficult to manage and as the networks have several types of apparatus ranging from switches and routers to middleboxes such as network address translators, firewalls, intrusion detection systems and server load balancers [3]. SDN is transforming the way we design and managing networks. The most important issue of SDN is Google's B4 data center networking system, which has taken 3 years to attain B4 propagation data [4]. B4 data center consolidation the traffic engineering works that govern links to nearby 100% utilization [5]. The principal benefits of SDN are unified network provisioning, superfluous pulverized safety, shortened outlay, lower managing cost, hardware savings and holistic enthusiasm governance [6], service provisioning speed and agility [7]. The brain of SDN is the controller, which consists of a set of applications that provide consolidated control functionality through open application program interface (API) to hoist the network forwarding behavior through an open interface [8]. The four communication interfaces are southbound,

M. R. Haque · S. C. Tan (✉) · S. Ali
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
e-mail: sctan1@mmu.edu.my; sameer.ali@szabist.edu.pk

C. K. Lee · Z. Yusoff
Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia
e-mail: cklee@mmu.edu.my; zulfadzli.yusoff@mmu.edu.my

I. R. Kaspin · S. R. Ziri
Telekom Malaysia Research and Development, TM Innovation Centre, Cyberjaya, Malaysia
e-mail: rizaludin@tmrnd.com.my; salvatore@tmrnd.com.my

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_11

northbound, eastbound and westbound. Controller communicates to application layer using northbound but to infrastructure layer via southbound. Eastbound and westbound interface are employed to communicate between controllers.

Distributed Denial of Service (DDoS) attack is an attempt to make an online service or network or SDN inaccessible by overburdening it with huge traffic from multiple sources around the world. PC, server, smartphone, alarm system, camera or any Internet-connected device can be the source of a DDoS attack. It can easily be created by sending hundreds of botnets that generate huge amount of traffic. It is very difficult to monitor DDoS onslaughts. In order to commence a workable DDoS attack, a DDoS attack can be generated by cyber aggressors who mostly form a network of computers or Internet-connected devices armed with a botnet [9]. Currently, one can purchase 10,000 botnets for $1000 [9]. It may become even cheaper in the future. Ericsson estimates that by 2020 there will be 50 billion devices interconnected worldwide [10]. These devices will probably be the potent parts of modern botnets and DDoS attacks. Attackers can enlist many machines because many machines are vulnerable without network security. Smart attackers leave no clue of their identity on the machines intruded. At present, no mechanism or tools are available to detect the source attacker.

Figure 1 shows an attacker initially sent a botnet to several devices around the world. Normally, user will not be able to realize that their devices have downloaded botnet. The botnet armies will then receive a command from the attacker to attack at predetermined times and generate huge amount of traffic to the victim's server or service of SDN application layer, data layer and control layer. Subsequently, the legitimate user's service will be interrupted. The magnitude and frequency of DDoS attacks are astonishing as aggressors make use of botnets to spread quickly due to modern-day telecommunication network interconnectivity. Spamhaus, an association that compiled lists of spammers, came under a voluminous Domain

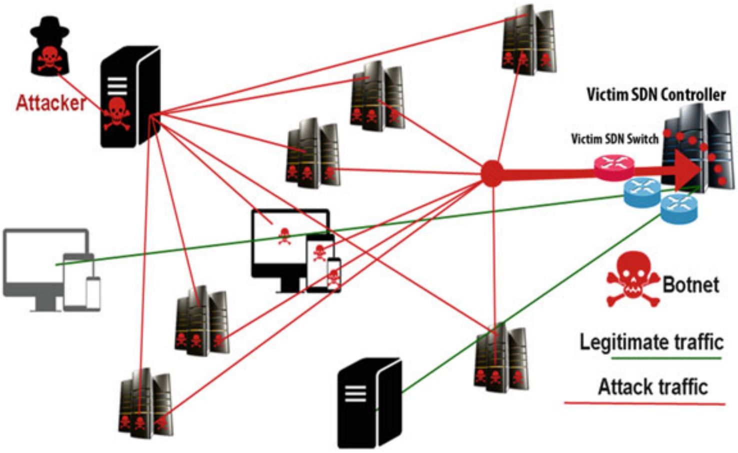


Fig. 1 DDoS attack on SDN using botnet

Name System (DNS) resemblance DDoS attack. In March 2013, the packed attack traffic was higher than 300 Gbps [11]. Good features of SDN may apply to reduce or avoid DDoS attack. Qiao Yan and F. Richard Yu in [12] propose some excellent features of SDN in reducing the impact of DDoS attacks by separating the control layer and data layer, progressive updating, software-based traffic analysis, programmable networking system and so on.

Since attackers associated each other to carry through effectual attacks, it is necessary for network providers to form an alliance to counter DDoS attack [9]. In SDN, controller is the main defender and has to collaborate with other controllers or layers to reduce or fend off DDoS attacks successfully. Authors in [13, 14] reported DDoS attack can flood controller-switch communication easily. In SDN, controller is the main defender since controller layer is the one making the decision to operate the SDN. It has to collaborate with other controller and layers to reduce or fend off attack effectively. Hence, it is essential to provide controller placement in SDN under DDoS attack. In this chapter, we will analyze the effectiveness of DDoS attack aware controller placement based on number of controller, switches, distance of links, data packets and bandwidth.

The flow of this chapter is as follows: Section 2 describes related work of SDN controller placement and DDoS attacks. Section 3 illustrates the DDoS attack propagations in SDN. Section 4 presents some models in DDoS attack aware controller placement based on hypothetical network in Malaysia where it can provide uninterrupted SDN services. Experimental results and analysis are given in Sect. 6.

2 Related Work

Various researches on controller placement in SDN have been reported. A. Tootoonchian et al. proposed greedy approach placement of controller algorithm by optimizing the reliability of SDN network [15]. B. Heller et al. suggested a radical rendering of SDN controller placement based on the latency of nodes to their controller. They tackled the issues of where and how many controllers to supply [16]. The authors in [17] used the k-median and the k-center algorithms under average-case latency and worst-case latency within the network elements. M. F. Bari et al. proposed an improved solution by dynamically provisioning controller. It is a framework that automatically adapts the number of controllers and links that are in service. [18]. Hu Yan-nan et al. developed various placement algorithms to establish possessed placement in order to maximize the reliability of SDN, since network disruption could effortlessly cause disconnections between the forwarding planes and control plane [19]. H. Aoki et al. discussed and formulated the controller placement problem in multi-domain networks on the basis of network partitioning. They commonly defined metrics of the controller placement, such as survivability and controller-switch latency that can be major on the separated networks [20]. Qiao Yan and F. Richard Yu highlighted that well-designed SDN configuration has ability to fend off DDoS attacks in cloud computing circumstance. The excellent features of SDN to counter DDoS attacks were summarized [12]. They also investigated that the new tendency and prominence of DDoS attacks in cloud computing, and provide an extensive assessment of fence mechanisms versus DDoS attacks using SDN [21]. Lim S. et al. proposed a DDoS obstructive application that enhances the popular SDN controller, POX [22]. However this application in only applicable to web server to safeguard DDoS attacks [23]. Salman, O. et al. [24] compared different controllers and their efficiency in performing the tasks. The presented results show that controllers using C and Java code have relatively quicker response as compared to controllers using python code. The performance, scalability, security and reliability among the different controllers in SDN were measured. E. Borcoci et al. [25] proposed a more accurate model to cater to various scenarios by extending the metrics size. Their findings also identified many uncovered research subjects adhering to static or dynamic work of the network system forwarding nodes to controllers, particularly when network links or nodes and/or controllers malfunction. B. Wang et al. reported a DDoS attack subsidence architecture that integrates an immensely programmable network governance to enable attack detection [26]. However, the above-reported works only addressed the controller reliability, latency and detection. Here, we proposed a model to provide uninterrupted SDN service under DDoS attack by the deployment of controller and backup controller placement.

3 DDoS Attack Propagation in SDN

In this section, DDoS attack propagation in SDN is illustrated. Open Networking Foundation (ONF) has taken the lead in defining SDN architecture [27]. The OpenFlow® protocol is a foundational component for making software-defined network fruitful. DDoS attack can propagate vertically and horizontally since SDN communicates one layer to another layer vertically and within the layer controller-to-controller horizontally [9].

Marc M. et al. [28] illustrated a typical propagation scenario between attacker and controller, shown in Fig. 2. Initially, DDoS attacker sends a huge traffic (defined as DDoS Attack Packets—DAP) to a switch and subsequently it will propagates to controller. The affected controller will then send to another switch upon request from the user. The user can be a legitimate user or attacker. This routine will continue until the affected controller becomes heavily congested due to excessive requests from the attacker and the service will then come to a halt. The backup controller will then be required to ensure the serve is not interrupted.

Qiao Yan [9] demonstrated the three possible propagation scenarios where attackers send DDoS attack packet (DAP) to the infrastructure layer in SDN architecture. Routers, switches, virtual switches and wireless access points receive traffic from the attacker, as depicted in Fig. 3. DDoS attacker can send thousands of DAP to any layer of SDN.

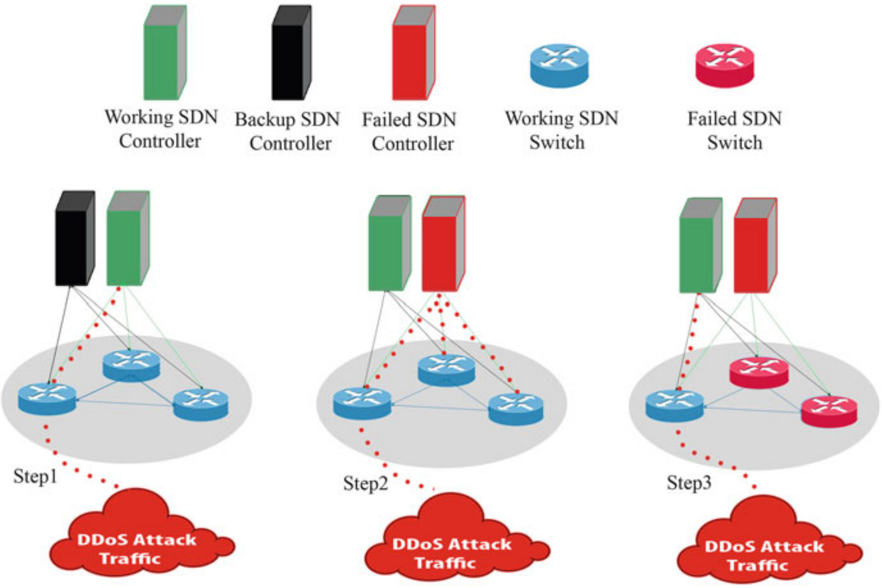


Fig. 2 Vertical propagation of DDoS attack on SDN switches and controller

Scenario 1:
In Fig. 3. DDoS attacker can send a huge amount of traffic on any application in application layer. Subsequently, this traffic may propagate to controller via northbound interface. This causes the affected controller to slow down and eventually the service come to a halt. This destruction will propagate down to the entire SDN architecture via southbound, eastbound and westbound.

Scenario 2:
Similarly, the attacker may propagate to infrastructure layer through northbound and southbound interfaces. Infrastructure layer is the propagation breeding area of DDoS attack in SDN [9]. In addition, the attackers can impulse an illustrious deal of the fake packets into SDN network as reported by [5].

Scenario 3:
Similarly, the attacker may propagate between switches or switch to controller through southbound interfaces.

In summary, it can be observed that DAP can propagate to any layer of SDN architecture. Switch and controller can fail to serve due to DDoS attack traffic. As a result, the SDN service will be interrupted by the attack. Here we are going to propose additional backup controller placement to ensure uninterrupted SDN services for legitimate user if the network is under attack.

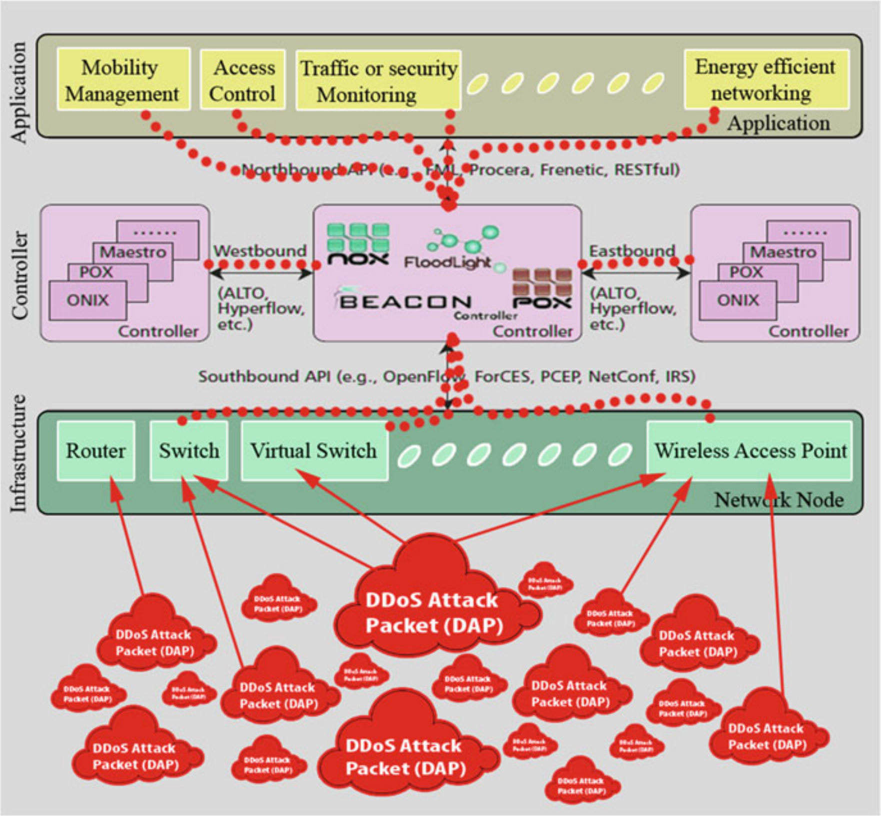


Fig. 3 DDoS attack on SDN architecture

4 Analysis of SDN Controller Placement Model in Malaysia

Here, we propose some models that can reduce the impact due to DDoS attack using different controller placement in different locations in SDN. In our study, we will evaluate the impact of number of controllers to be deployed under three different hypothetical network scenarios of controller placement across Malaysia.

At first we placed one controller in one location (either Zone 1 or Zone 3) connected to all switches to be shared for the whole country to investigate the impact of distant between the southern and northern or vice-versa. Secondly, we connected one controller at the center of Malaysia, which enabled us to observe the impact of equal distannce between the southern and northern. Finally, we connected one controller in every location, namely, Zone 1, Zone 2 and Zone 3 with backup links.

Scenario 1:
Figure 4 shows one SDN controller installed in Johor (Zone 3, southern), which connected to 10 SDN switches to serve the three zones under this SDN. From the

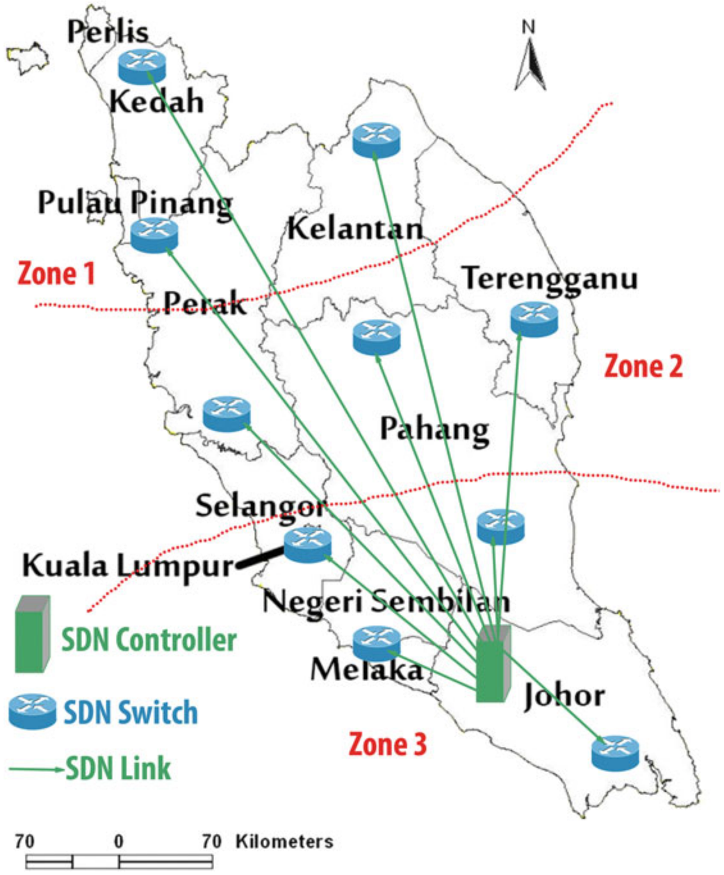


Fig. 4 One SDN controller placement at Johor with 10 switches

results obtained, reliability, scalability, latency and propagation delay performance of Zone 3. However, relatively poorer overall performances are observed in Zones 1 and 2. This is due to the impact of distance.

The approximate distance from controller at Zone 3 (southern) to Zone 2 (center) and Zone 1 (northern) are 300 km and 600 km, respectively. The radius of Zone 3 is approximately 150 km. The controller can perform only in this zone with minimum cost of distance and availability. The DDoS attack risk is high in this model because there are no backup controllers here to serve during DDoS attack.

Scenario 2:
The second model is created where one controller connected with 10 switches from the center of Malaysia, as shown in Fig. 5. Compared to the first model, the second model is less cost of distance for the link around controller but the DDoS attack risk is still high in this model as this model also doesn't have any backup controller.

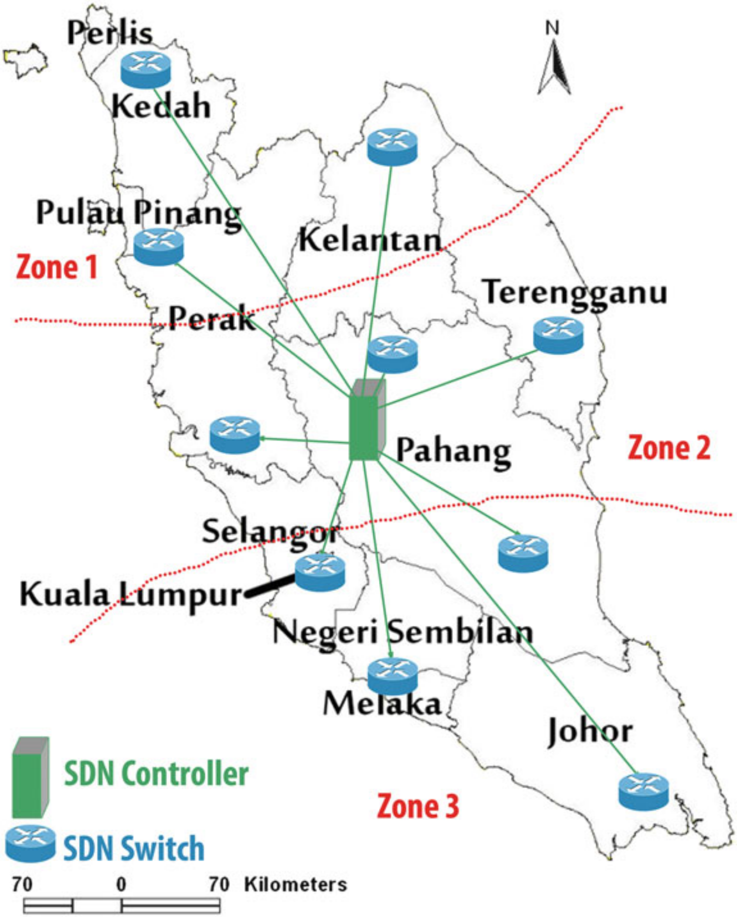


Fig. 5 One SDN controller placement at Pahang with 10 switches

In Fig. 5, we observed that one SDN controller has been installed at the center point of west Malaysia in Pahang. It is much better in terms of cost and performance compared to Johor (Zone 3, southern). But the best performance of controller is in zone 2 (Pahang, center). Zone 1 (northern) and Zone 3 (southern) have similar performance in terms of cost and reliability.

Scenario 3:
Figure 6 shows three SDN controllers installed in three different places in Johor (southern), Pahang (center) and Pulau Pinang (northern). In this model, we placed three controllers and these controllers can serve strongly 150 km around them. Similar performances of each controller are obtained in all three zones. The main benefit of this model is that if any controller becomes the victim of DDoS attacker, the nearby controller can easily be alerted from infected controller if that particular

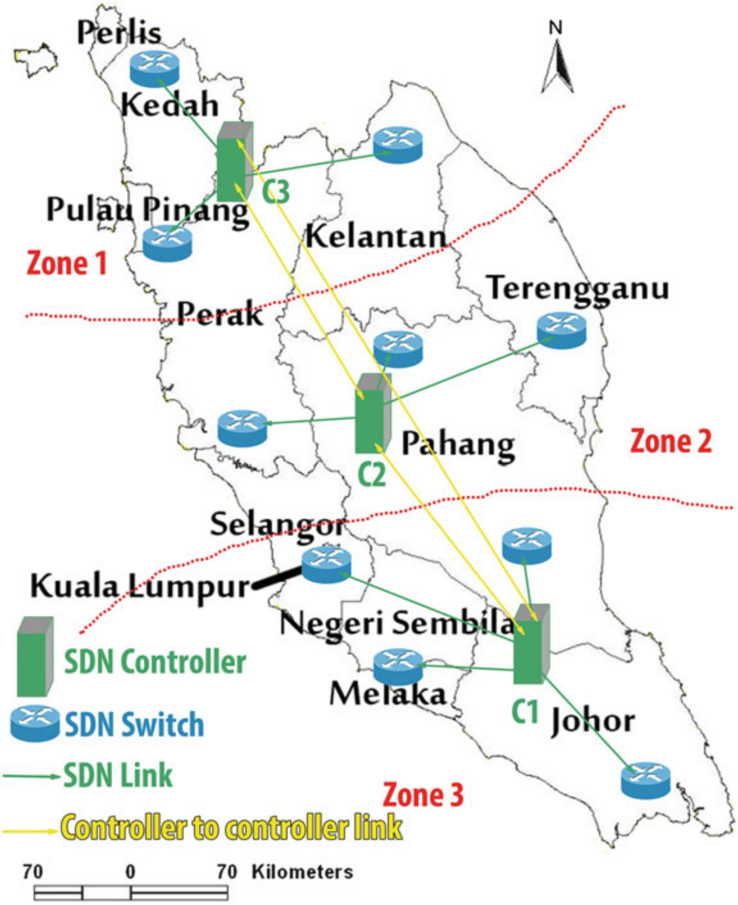


Fig. 6 Three SDN controller placement with 10 switches

controller is being attacked. When the controller of Johor (Zone 3, southern) under attack, the controller installed at Pahang (Zone 2, center) will be called upon to serve as a backup controller. Similarly, this Zone 2 (center) controller can be utilized if Zone 1 (northern) is under attacked.

Now, the effectiveness of our proposed model is tested with the following conditions. Figure 7 illustrates where Zone 1 and Zone 3 are under attack.

The performances are tabulated in Table 1.

From the table, it is evident that one controller placement has higher chances of DDoS attack compared to three controller placements; however, the cost incurred is high. Further tests conducted through a simulation of hypothetical network will be presented in following.

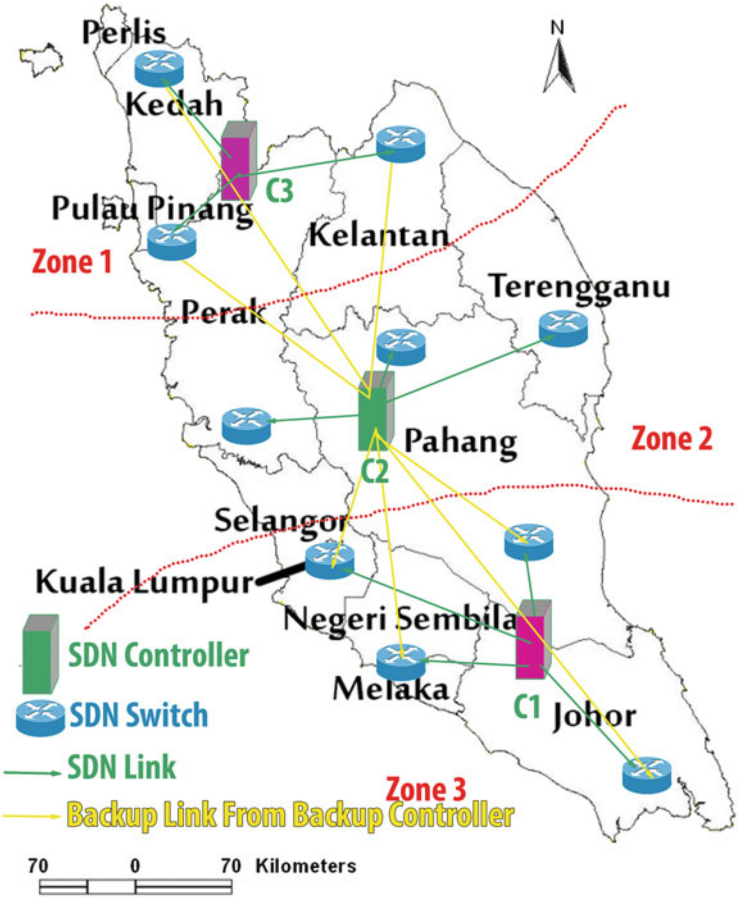


Fig. 7 One nearby backup SDN controller placement and two failed controllers with 10 switches

Table 1 The SDN controller placement performance table

Number of controller placements on SDN	Number of switches on SDN	Best zone of SDN	Placement cost	Chance of DDoS attack
1	10	Zone 3	Low	High
1	10	Zone 2	Low	High
3	10–30	Zone 1,Zone 2,Zone 3	High	Low

5 Experimental Results from Quantitative Value

In this section, we will demonstrate the operational of controller placement under DDoS attack based on quantitative value of controller, switches, distance of links, data packets and bandwidth.

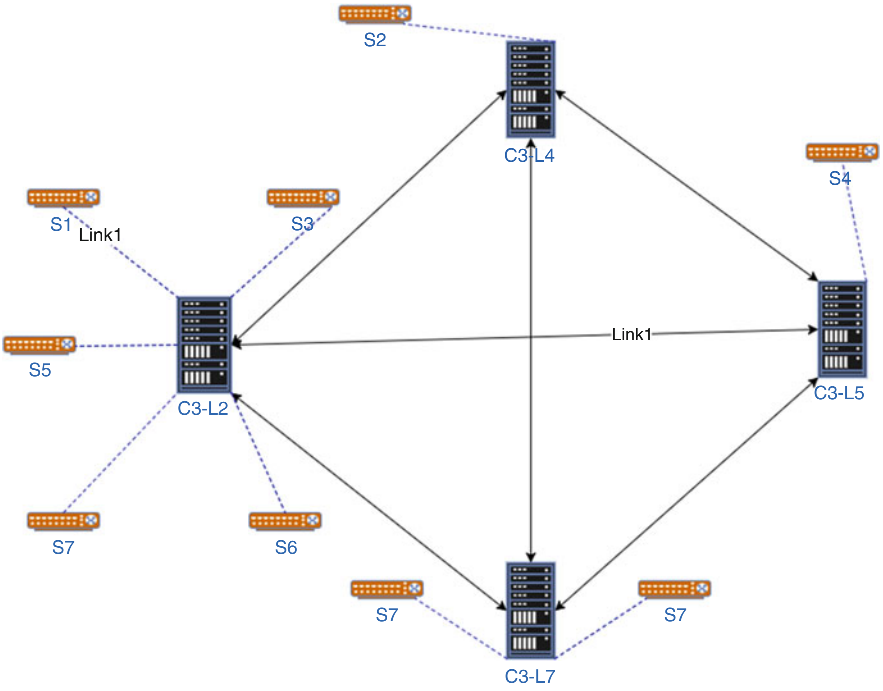


Fig. 8 Four SDN controller placements with nine switches without DDoS attack

Analysis 1:

In this scenario, to optimize the total cost, four C3 type controllers with the processing power of 8000 packets per second (pps) costing $4500 with 64 available ports each have been selected by the model to process the total available data packets 28,840 pps. To connect controller to controller, controller to backup controller and controller to switch it selected L1, which is the lowest cost at $0.25 and 1,00,00,000 bps bandwidth. No backup controllers have been placed as there are no DDoS attacks on any location (Fig. 8).

Analysis 2:

In this scenario, to optimize the total cost, one C1 and one C3 type controller with the processing power of 8000 pps costing $4500 with 64 available ports each have been selected by the model to process the total available data packets 8800 pps from three switches. To connect controller to controller, controller to backup controller and controller to switch it selected L1, which is the lowest cost at $0.25 and 1,00,00,000 bps bandwidth. One backup controller BC1 has been placed as there are DDoS attacks on the location L8 (Fig. 9).

Analysis 3:

In this scenario, to optimize the total cost, one C3, two C6, four C9 and one C13 type controller with the processing power of 8000 pps, 13,000 pps, 13,000 pps and 13,000 pps, costing $4500, $9500, $9500 and $9500, respectively, with 64, 128,

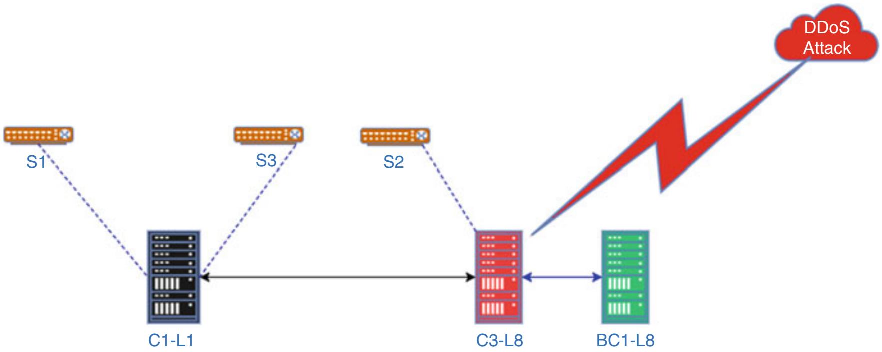


Fig. 9 Two SDN controller placements with three switches with one backup controller under DDoS attack

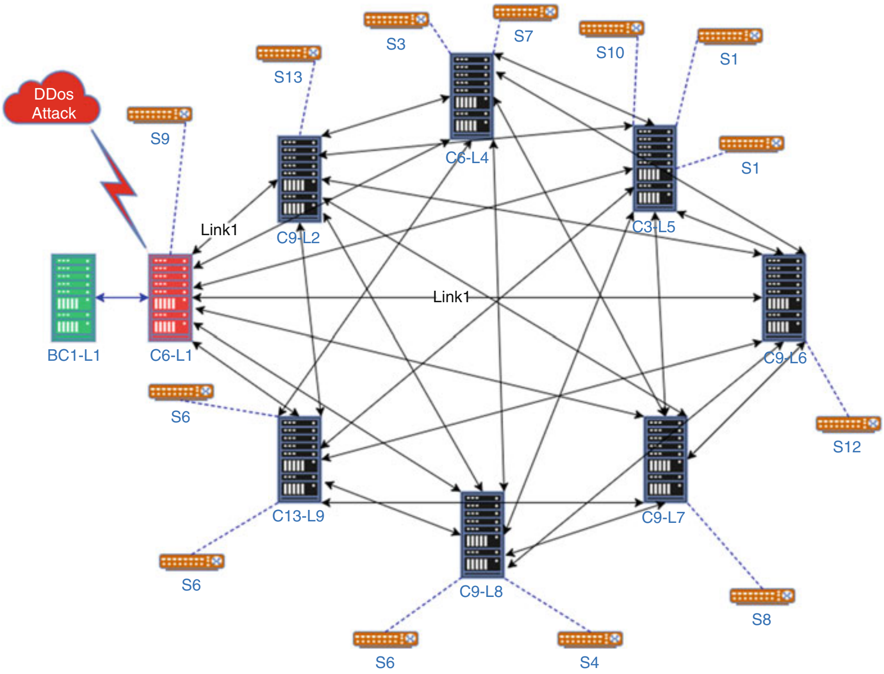


Fig. 10 Eight SDN controller placements with 13 switches and 1 backup controller under DDoS attack

128 and 128 available ports each has been selected by the model to process the total available data packets 71,900 pps from 13 switches. To connect controller to controller, controller to backup controller and controller to switch it selected L1, which is the lowest cost at $0.25 and 1,00,00,000 bps bandwidth. One backup controller BC1 has been placed as there are DDoS attacks on the location L1 (Fig. 10).

6 Conclusion and Discussion

In this chapter, the details of DDoS attack propagation in SDN are presented. Our proposed DDoS attack aware controller placement was tested initially for a hypothetical network in Malaysia and later extended to a real practical network successfully.

We have investigated different previous research works done on SDN in the context of DDoS attacks; most of the work focused on the placement of controllers, determining number of controllers, issues of reliability, scalability, latency and propagation delay. Finally, we showed some new hypothetical models of controller placement in different places to reduce DDoS attacks in SDN. We observed that one controller placement is not sufficient to reduce DDoS attack and the whole SDN will malfunction as a result. More than one controller is needed to be positioned in various places to reduce the chances of malfunction of whole SDN due to DDoS attack.

The outcome of this work is essential in SDN to provide continuous services.

Acknowledgment This research work is fully supported by Telekom Malaysia (TM) R&D and Multimedia University (MMU), Cyberjaya, Malaysia. We are very thankful to the team of TM R&D for providing support to our research studies.

References

1. P., Xia., L. Zhi-yang, G. Song, Q. Heng, Q. Wen-yu, Y. Hai-sheng. AKself- adaptive SDNcontroller placement for wide area networks, Frontiers of Information Technology & Electronic Engineering, ISSN 2095-9184 (print); ISSN 2095-9230 (online). 2016.
2. M. Seliuchenko, Orest Lavriv, Oleksiy Panchenko, Volodymyr Pashkevych, Enhanced Multicommodity Flow Model for QoS-aware Routing in SDN, 2016 International Conference “Radio Electronics & InfoCommunications” (UkrMiCo) September 11–16, 2016, Kiev, Ukraine.
3. N. Feamster, J. Rexford and E. Zegura, The Road to SDN: An Intellectual History of Programmable Networks, 16-Jan-2016, Available: https://people.csail.mit.edu/alizadeh/courses/6.888/papers/sdnhistory.pdf.
4. Jain, S., Kumar, A., Mandal, S., Ong, J., Poutievski, L., Singh, A., Venkata, S., Wanderer, J., Zhou, J., Zhu, M., et al. (2013). B4: Experience with a globally-deployed software defined wan. In ACM SIGCOMM Computer Communication Review, volume 43, pages 3–14. ACM.
5. Q. Yan, Qingxiang Gong and Fang-An Deng, Detection Of Ddos Attacks Againstwireless Sdn Controllers Based On The Fuzzy Synthetic Evaluation Decision-Making Model, Ad Hoc & Sensor Wireless Networks, Vol. 33, pp. 275–299, September 12, 2016.
6. G. Brown, 7 Advantages of Software Defined Networking, Ingram Micro Advisor, 8/12/2014, Available: http://www.ingrammicroadvisor.com/data-center/7-advantages-of-software-defined-networkingR. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
7. S. Yegulalp, Five SDN Benefits Enterprises Should Consider, Newtwork Computing, 2013. Available: http://www.networkcomputing.com/networking/five-sdn-benefits-enterprises-should-consider/70381323.

8. M. D. Yosr Jarraya and T. Madi, “A survey and a layered taxonomy of software-defined networking,” IEEE Commun. Surveys Tuts., vol. 16, no. 4, pp. 1955–1980, 4th Quart. 2014.
9. Q. Yan, F. Richard Yu, Senior Member, IEEE, Qingxiang Gong, and Jianqiang Li, Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 1, FIRST QUARTER 2016
10. Ericsson, Heterogeneous Network (Hetnet), https://www.ericsson.com/br/res/thecompany/docs/press/media_kits/hetnet_infographic_vertical_04.pdf, 2017.
11. “Arbor special report: Worldwide infrastructure security report volume IX,” Arbor Netw., Inc., Burlington, MA, USA, Tech. Rep., [Online]. Available: http://pages.arbornetworks.com/rs/arbor/images/WISR2012EN.pdf
12. Q. Yan and F. Richard Yu, Security And Privacy In Emerging Networks, Distributed Denial of Service Attacks in Software-Defined Networking with Cloud Computing, IEEE Communications Magazine, April 2015.
13. D. Kreutz, Fernando M. V. Ramos, Paulo Verissimo, Christian Esteve Rothenberg, Siamak Azodolmolky, and Steve Uhlig, Software-Defined Networking: A Comprehensive Survey, arXiv:1406.0440v3 [cs.NI] 8 Oct 2014.
14. S. Scott-Hayward, SDN Security: A Survey, SDN4FNS – November 2013.
15. A. Tootoonchian S. Gorbunov Y. Ganjali M. Casado and R. Sherwood, in Proceedings of the 2nd USENIX conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services, ser. Hot-ICE’12. Berkeley, CA, USA: USENIX Association, 2012, pp. 10–10.
16. B. Heller, R. Sherwood, and N. McKeown, in Proceedings of the first workshop on Hot topics in software defined networks. ACM, 2012, pp. 7–12.
17. S. T. Zargar, J. Joshi, and D. Tipper, “A Survey of Defense Mechanisms Against Distributed Denial of Service (DDoS) Flooding Attacks,” IEEE Commun. Surveys & Tutorials, vol. 15, no. 4, 2013, pp. 2046–69.
18. M. F. Bari et al., Proceedings of the 9th International Conference on Network and Service Management (CNSM 2013), Zurich, 2013, pp. 18–25.
19. H. Yan-nan, WANG Wen-dong, GONG Xiang-yang, QUE Xi-rong, CHENG Shi-duan. Beijing University of Posts and Telecommunications, Beijing 100876, China, The Journal of China Universities of Posts and Telecommunications, 2012.
20. H. Aoki, Norihiko Shinomiya, Controller Placement Problem to Enhance Performance in Multi-domain SDN Networks,Graduate School of Engineering, Soka University, Tokyo, Japan The Fifteenth International Conference on Networks (includes SOFTNETWORKING, 2016).
21. Q. Yan, F. Richard Yu, Senior Member, IEEE, Qingxiang Gong, and Jianqiang Li, IEEE Communications Surveys & Tutorials, Vol. 18, No. 1, First Quarter 2016.
22. M. Cauley, M. (2013). About POX. URL: http://www.noxrepo.org/pox/about-pox/. Online.
23. Lim, S., Ha, J., Kim, H., Kim, Y., and Yang, S. (July 2014). A SDN-oriented DDoS blocking scheme for botnet-based attacks. In 2014 Sixth International Conference on Ubiquitous and Future Networks (ICUFN), pages 63–68.
24. S., Ola., Elhajj, I.H., Kayssi, A. and Chehab, A., 2016, April. SDN controllers: A comparative study. In 2016 18th Mediterranean Electrotechnical Conference (MELECON) (pp. 1–6). IEEE.
25. E. Borcoci, Radu Badea, Serban Georgica Obreja, Marius Vochin, On Multi-controller Placement Optimization in Software Defined Networking – based WANs” University POLITEHNICA of Bucharest – Romania, IARIA, ISBN: 978-1-61208-398-8, 2015.
26. B. Wang, Yao Zheng, Wenjing Lou, Y. Thomas Hou DDoS attack protection in the era of cloud computing and Software-Defined Networking,Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, Computer Networks 81 (2015) 308–319, 2015.
27. Open Networking Foundation, Software-Defined Networking (SDN) Definition, 2017, Available: https://www.opennetworking.org/sdn-resources/sdn-definition.
28. M. Manzano, Anna Manolova Fagertun, Sarah Ruepp, Eusebi Calle, Caterina Scoglio, Ali Sydney, Antonio de la Oliva, and Alfonso Mu ˜noz, Unveiling Potential Failure Propagation Scenarios in Core Transport Networks, arXiv:1402.2680v1 [cs.NI] 11 Feb 2014

Part III
Data Access and Visualisation: Tools and Platforms/Systems

Multimodal Systems, Experiences, and Communications: A Review Toward the Tactile Internet Vision

Mohammad Al Jaafreh, Majed Alowaidi, Hussein Al Osman, and Abdulmotaleb El Saddik

1 Introduction

In the past decade, multimedia has been widely based on vision and hearing. However, many forms of content especially those that stimulate other senses including olfaction, tactile, kinesthetic, and gustatory have been recently proposed [1–3]. Therefore, we are witnessing a rapid evolution of new technologies that extends the audiovisual services into multimodal (i.e., multisensory) applications. One of the promising technologies is the Tactile Internet [4]. The Tactile Internet (5G) is a huge paradigm shift ahead [5]. It enables users to engage with objects in a local/remote environment bidirectionally and enjoy the offered ultra-low delay communication services.

> Multimodal communications is the field referring to the representation, storage, retrieval, and dissemination of temporally, spatially, and contextually correlated multisensory information expressed in multiple media such as text, voice, graphics, images, animations, audio, video, smell, and touch.

These new modalities have been proven to potentially enrich the overall users' experience and satisfaction in many applications ranging from interactive gaming to medical and military training [6–8]. In the following paragraphs, we provide a brief description of the haptic, gustatory, and olfaction modalities.

M. Al Jaafreh (✉) · M. Alowaidi · H. Al Osman · A. El Saddik
EESC, University of Ottawa, Ottawa, Canada
e-mail: jaafreh@uottawa.ca; malow039@uottawa.ca; halosman@uottawa.ca; elsaddik@uottawa.ca

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_12

Haptic is referred to as the science of touch. The term is derived from the Greek word *hapt esthai* that refers to the sense of touch [2, 9]. The haptic perception was examined by many studies [2, 10, 11]. A haptic process can be considered as force, tactile vibration, or heat stimuli that represent mechanical, thermal, or pain receptor. Furthermore, haptic manipulation can be achieved in one of the following forms: tactile feedback, kinesthetic feedback, or motor action. A haptic signal has a unique property of being both informatic and energetic [12, 13]. It is characterized by a bidirectional flow of information. Tele-haptics extends the haptic capability to include remote distance interactions [2]. The haptic interfacing device has to provide all the possible degrees of freedom (DoF) inherited from the biomechanical structure of humans' arms, joints, hands, etc. Nowadays, Sensible Phantom Omni [14] device provides six DoFs positional sensing and three DoFs of force feedback, which allow the human nervous system to interact haptically with object in a very effective manner. As such, augmenting virtual environment and/or telepresence with haptic features implies that, *togetherness*, *tele-manipulation*, *and tele-touching* can be achieved among users in a realistic manner.

Olfaction is referred to as the act of smelling [15–17]. More specifically, it is the nose's ability to perceive a scent of material or substance in an environment. In spite of the fact that the olfaction modality can enrich the multimedia system, this field has been nearly neglected due to the limitations of hardware manufacturing required to remotely produce the desired scents [18]. However, many successful endeavors have lately succeeded in enhancing the quality of scent displays [19]. Furthermore, [15, 20, 21] conducted studies on olfactory displays and smell sensors to show the potential of transmitting scent information such as aroma of fruits, cooked food, and flowers to users over networked virtual environments. In addition, the authors of [19] investigated the current use of olfaction in many multimedia fields, from the filming industry to Virtual Reality (VR) games [20].

The gustatory sense represents the taste sensation that can be perceived from a mixture of chemical signals [3, 16]. Gustation sensation involves a flavor of multifaceted or a mixture of taste sense. Flavor is defined by ISO [22] as different sensory integration modals of olfaction, gustatory, and trigminal that are recognized during the taste process. Generally, taste is categorized in five basic qualities: sweet, sour, salty, bitter, and umami. Due to the complexity of chemical combinations required to produce distinct tastes and the fact that the gustatory sense is affected by other modalities like olfaction, vision or thermal, the research conducted on the perceived gustatory sense is still limited. Recently, [23] analyzed the diachronic and synchronic experiences of the five taste groups in order to establish a generic framework to design a taste experience human–computer interaction (HCI).

With the rapid development in the areas of multimedia hardware and with the emergence of the Tactile Internet 5G technology, traditional media systems can be evolved into collaborative, interactive, and immersive multimodal systems. The multimodal integration of audio, video, three-dimensional (3D) graphics, olfaction, haptic, and gustatory modalities allows users to not only watch and listen to information from remote environments, but it also provides a high-quality immersive interaction in which, they can collaboratively touch, manipulate, and smell objects from remote environments [24–27]. For instance, in addition to benefiting from high frame rate, dynamic range, and ultra-high-definition videos in multimedia applications, users now can physically feel, smell, and taste computer-simulated objects. This in fact creates a unified perception that boosts people's experience. The potential uses and advantages of multimodal media systems are discussed in several studies. In [28], authors present a comprehensive review covering multimodal fusion methods that are focused on multimedia analysis fields. They categorized multimodal fusion methods to rule-based, classified-based, and estimation-based methods. Although they focused on fusing audio, video, and text modalities, the work did not address new surfacing modalities that are necessary nowadays in modern multimedia systems such as haptic, olfaction, gustatory, etc. In [29], the author stressed the requirements for tourism experiences to move beyond the audio-visual domain. Therefore, she investigated the mechanisms necessary to facilitate multisensory tourism experiences. More specifically, the author highlighted how important the senses of sight, sound, taste, touch, and smell are to the tourist experience of the Canadian Great Bear Rainforest. Also, multimodal systems have been used for therapeutic purposes. Providing multisensory media [30] channels of medication like aroma therapy, tactile stimuli, and mellowing music could provide a robust experience to impaired people with Alzheimer's, autism, and dementia [31]. Another project called Multisensory Systems [32] presents systems that provide multiple sensory simulations comprising of olfaction, visual imagery, vibration, and 3D sound that have been proven to enhance the overall Quality of Life (QoL) [30]. Authors in [33] designed and implemented a gustatory-based game called LOLLio. In the game, a haptic input device is used as an interactive lollipop capable of contextually producing two taste qualities (sweet and sour). The game was evaluated using early age participants where LOLLio was served as a movement controller in a game like Pacman. If the child achieves the game goal, he or she will be rewarded by a pleasant sweet taste, whereas losing the game challenge can cost the player a punishment of sour stimuli.

The advantages of local multimodal media systems and how user perceives this content have been discussed in several studies [6–8, 30]. Conversely, the communication of multimodal information over the Internet is still an open research question. A suitable communication protocol for such information has to adapt and/or meet the user, application, modality, and network requirements. Those requirements have to be taken into consideration when defining the schema of the new Tactile Internet 5G infrastructure. In fact, interactivity, collaboration, co-presence, and togetherness cannot be achieved without tackling the communication part. Therefore, surveying the existing communication frameworks that are capable of disseminating the information generated by multimodal systems without adversely affecting the rendering quality of the received content is very crucial. In this chapter, we review the most in-depth listing of the technical methods required to realize a suitable multimodal communications protocol. Section 2 presents the QoS requirements for each modality. A holistic study of QoE in interactive multimodal systems is presented in Sect. 3. Section 4 discusses the most in-depth technical methods required to realize a suitable multimodal communication protocol. Section 5 provides a survey of the state of the art of communications protocols. The review complemented the discussion based on the related works in this field, specifying the advantages and disadvantages of existing protocols. Section 6 provides the ultimate endeavor to realize our vision on a new paradigm referred to as the Internet of Multimodal Things (IoMT) over Tactile Internet. The chapter is concluded in Sect. 6, providing further future research directions.

2 QoS Requirements for Each Modality

Traditionally, before the rapid emergence of the vast amount of multimedia content on the Internet, computer networks were designed to convey data traffic using a scheme known as best-effort delivery. This scheme was somewhat acceptable in the past, where data traffic was elastic (i.e., it can stretch under delay and bandwidth impairments). This best-effort service is unable to provide a predictable and reliable end-to-end packet delivery for real-time services and business mission-critical applications. The recent emergence of newly perceived multimedia classes of applications has highlighted the need for the development of new standards that can accommodate their characteristics. This has been recently realized in the form of a delivery scheme known as Quality of Service (QoS).

> QoS refers to a set of quality metrics used to tune and quantify the performance of applications, systems, and networks.

From a network point of view, QoS is defined as the set of service requirements that have to be met to enhance the overall utility of the network [34, 35]. This can be done by granting priority to higher value or more performance-sensitive flows. Generally, QoS metrics are grouped into application factors, system factors, and network factors. Metrics in each group are bidirectionally affected by metrics in the other two groups. According to several research studies [36–41], the most significant factors of the QoS metrics are considered: packet size, delay, jitter, throughput, data loss rate, rendering (update) rate, and arrival model.

- **Packet type and size**: Define how the modality is encoded and transmitted over the Internet. The size of the packet is defined based on the codec standard.
- **Latency/delay**: The round-trip time for the packet to be sent over the network from source to destination and back. It is usually measured in millisecond or microsecond. Delay can be further categorized into propagation delay and transmission delay. The propagation delay is caused by the resistance of a physical medium. The transmission delay is determined by link bit rate, time spent in router queues, and number of hops in the selected route. In this context, we are considering the total end-to-end delay.

- **Jitter**: The difference in latency of network packets usually measured in microseconds or nanoseconds.
- **Bandwidth/throughput**: The amount of data transferred from source to destination or processed in a given amount of time. Measured typically in bits/second or bytes/second.
- **Data loss rate**: The percentage of data that have been sent and not delivered to the receiver.
- **Update rate**: How many times per second the data of each modality need to be refreshed. It is different than the sampling rate and is measured in Hz.
- **Arrival model**: Defines the arrival and losses characteristics of the modality packet in the Internet. It can be either Periodic, Markov Model, or Bernoulli Model.

The requirements for these five QoS factors in relation to the aforementioned media types are discussed in this section, and summarized in Table 1.

Internet-based haptic application enables the interaction between a human operator and a remote actuator. Such system is referred to as a Master–Slave Teleoperation (MST). MSTs are measured by the number of DoFs they provide [42] and can be used in several fields, such as military space robotics, underwater operations, and remote medical surgeries. The haptics' bidirectional data flow property implies a distinguished requirement for teleoperation systems, which maintains local control loop stability [43]. Unstable teleoperation interaction is a resultant of unwanted delay, jitter, and packet loss. Haptic data is more sensitive to delay since the operation performance of kinesthetic or tactile feedback is typically positioned between 1 and 50 ms. When the MST involves Machine to Machine (M2M), the latency requirement is extremely critical with a range of 1–10 ms, in order to interact with fast-moving objects [11, 44, 45]. In MST systems, we might witness two scenarios, the master–slave, the sent date can be among others force, torque, or velocity (in three dimensions) sent in a 1 kHz rate, and the feedback from the slave, which can be in the form of verbotactile information or a kinesthetic with 1000 packet/s. The 100 DoF tactile is denoted to include most of the body's biomechanical structure. Unstable latency (jitter) has the most adverse effect on the teleoperation process in terms of stability of performance. To ensure stability and transparency for MST, jitter has to be less than 2 ms, and data loss has to be between 0.01% and 10% [43, 46]. To achieve a high-fidelity output, the update rate should be greater than 1 kHz. This means that the haptic device should receive 1000 frames per second to render the output in a good resolution [47]. The bandwidth is another requirement of teleoperation communication; however, it has less of an impact since kinesthetic, tactile, and actuators have small volume of data per frame (usually 512 kbps) [46]. The arrival model of the haptic modality is heterogeneous by nature and can be described using the Gilbert–Elliot model with two-state Markov process [48]. In summary, haptic systems are very critical in terms of delay, jitter, and update rate and much robust to data loss and bandwidth.

Video represents a sequence of continuous frames of visual data synchronized in a timely fashion [49, 50]. Video conferencing and streaming is another example

Table 1 Communication requirements for IoMT over Tactile Internet

QoS metric	Human-to-machine haptic	Video	Audio	3D	Olfaction	Machine -to- human haptic feedbacks	
Packet type	(Positions, Velocity, Force, Torque)	H.264/MPEG-4	Dolby Surround	Mesh, Texture	Scent	Kinesthetic Signals	Tactile Signals
Packet size (B)	1 DoF: 2–8 3 DoFs: 6–24 6 DoFs: 12–48	1.5K	>50	1.5	1	1 DoF: 2–8 3 DoFs: 6–24 6 DoFs: 12–48	1 DoF: 2–8 10 DoFs: 20–80 100 DoFs: 200–800
Jitter (ms)	1–2	30	30	30	≤23	2	1
Delay (ms)	1–50	≤400	≤150	100–300	0–1500	10	1
Throughput (kb/s)	≥ 512	≥2500	≥128	≥1200	0.008	≥ 512	≥ 1000
Data loss rate (%)	0.01–10	1	1	1–10	1	10	0.01
Update rate (Hz)	≥1000	30	20	30	0.1–10	≥ 500	≥ 1000
Arrival model	Heterogeneous (Periodic or Gilbert–Elliot)	Periodic	Periodic	Periodic	Periodic	Heterogeneous (Periodic or Gilbert–Elliot)	Heterogeneous (Periodic or Gilbert–Elliot)

of an application that cannot survive on a best-effort standard for routing through an Internet Protocol (IP) network [51]. One point worth noting is that in terms of video communication and compression of the MPEG-V standard, only three types of frames are taken into consideration [49, 50]: Intra-coded frames (I-frames) are used for coding real-time images frames that do not have references of previous frames, Predictive-coded frames (P-frames) are used to reference the changes in previous I-frames or P-frames, and Bidirectional-predictive frames (B-frames) are used as data reference for previous and successive frames, so B-frames are aimed to enhance quality of compression. To prevent video flickering, frames have to be refreshed momentarily [52]. The refresh rate of video varies from one application to another; however, 30 frames per second is the common rate. In video applications, QoS communication requirements should be preserved within rigged time-based parameters in order to maintain temporal relation between information entities [35, 46]. A delay of 400 ms or less should be preserved to provide an acceptable human perception of real-time scenes. The jitter should be kept below 30 ms. Real-time video applications require a large volume of data per frame; therefore, a high bandwidth provision may vary from 2.5 to 5 Mbps. In order to render a rich video application, the data loss should be constrained to around 1% and the update rate has to be around 30 frames per second [8]. Due to the manifested usage demands of high-resolution video content, higher dynamic range and higher frame refresh rate (60 fps) are significant to optimize video delivery [53].

Audio is defined as a continuous waveform of perceived voice/sound signal that propagates in a medium [49]. Voice-over IP (VoIP) applications such as Skype, Viper, and Hangouts are getting more popular for social multimedia users. VoIP applications transfer voice packets over an IP network, while requiring guaranteed bandwidths and very low delay and jitter. Audio delays expectations are based on the interaction level of the application, that is, one-way, two-way, or asymmetric two-way [54]. Generally speaking, the preferred range of the delay is less than 150 ms, whereas the jitter should stay around the 30 ms boundary. The audible frequency threshold that can be perceived by the human ear is roughly 20 Hz [55], hence the update rate has to be at least 20 frames per second. It has been shown [56] that the tolerated data loss of audio applications should be preserved below 1% and the throughput has to be more than 200 Kbps.

3D graphics is the representation of three dimensions of data in a geometric space. This space can be a composite of dimensions selected from length, width, and depth. 3D plays a significant role in the creation and development of virtual reality (VR) and augmented reality (AR) applications [57]. VR is described as a computer software that simulates a real or virtual physical existence where people have the ability to interact with it instantaneously and change it as if it is real. Augmented reality is the technique of enhancing the real world by getting information virtually to enhance a user's experience and senses. Multiview and stereoscopic scenes can enhance 3D graphics, which in turn will improve humans' perception [53]. 3D streaming has been introduced to overcome the limitations of preinstalling and/or downloading the vast virtual environments (VE) content. 3D streaming is defined as continuous and real-time delivery of 3D contents, such as meshes, texture, and animations, over network connections to allow user interactions with its virtual world without a full download or a preinstallation [58–60]. Users can immediately render the 3D content when it is only partially received, and thus the interaction with their VE occurs without having to wait for the entire download to be completed. 3D streaming is similar to audiovisual streaming, where users can immediately interact with the displayed scene when the data are gradually downloaded. However, there exist some key differences between both approaches. The first difference relates to the content itself, which is 3D mesh models, texture, and animation for the 3D streaming and 2D images for audiovisual streaming. This means that the user in 3D streaming can navigate the same scene at its all-possible resolutions and from any viewing angle without the need to download any extra data. However, in audiovisual streaming when a user wants to change his/her viewing angle, extra images for the same objects in the scene have to be streamed. In audiovisual streaming, the video is fragmented into sequentially ordered frames. These frames are transmitted according to the time sequence of the video. Hence, the data stream would be the same for everyone. The audiovisual streaming access pattern is considered linear and frame prediction can be applied; whereas in 3D streaming, the 3D content cannot be ordered and has to be fragmented based on the user's behavior, viewing angle, and distance. Different viewing angles or distances thus would produce unique transmission sequences. Obviously, in 3D streaming the content access pattern is hard to be anticipated and lacks linearity [59]. The QoS factors for networked virtual environments have been evaluated in [61]. The findings are as follows: the end-to-end delay fluctuates between 100 ms and 300 ms, the jitter effect should be at most 30 ms, the 3D contents has to be periodically updated in a 30 frame per second rate, the data loss is equal or below 10%, and the bandwidth should be greater than 1.2 Mbps.

Tele-olfaction refers to the transmission of scents over a network [16, 62]. Users can sniff a mixture of odorants located at a remote place using an aromatic sensing system combined with olfactory display. Unlike other media types, there is a fundamental challenge before realizing scents on demand. Given the fact that thousands of kinds of olfactory receptors are found in the human nose [1], researchers find difficulties in creating a systematic, solid, and standard scheme for olfaction. Although this topic is still relatively understudied, a few works have been carried out to determine the tele-olfaction QoS metrics. In terms of virtual olfactory system, authors in [63] itemized the latency of any olfaction display in terms of the purging(cleaning) time for previous odor, the odorant forming time, and the time of delivery. An olfactory study [20] was conducted to assess the delay influence on a game when aromatic information is delivered to a player in a networked fruit-harvesting game. The study found that to get a correct smell judgment from the player, the maximum allowable jitter has to be less than 23 ms. Even though a human can indicate the smell source (in respect to median or sagittal plane) is about 0.1 ms [63], the perceived delay ranges between 0 and 1500 ms. Regarding the bandwidth needed to convey scents, a tele-olfactory system needs to rely on qualities of scents rather than their quantities, as indicated in [19], where some scents have different threshold than others. Also [63], after carrying out a test on 60 odor classes, each one in a four concentration scale, claimed that at most eight bits of capacity channel are needed to carry olfactory information. Beside the aforementioned QoS metrics for olfaction, [63] performed a study that investigates the required refresh rate for a virtual olfactory application needed over network. They found that the requirement for the update rate is 0.1–10 Hz and the shifting time between aromas is between 1 and 10 s.

The recent implementation of the gustation is based on the cross-modal interaction between VR and olfaction. Hence, gustatory QoS parameters are dependent on visual and olfaction communications factors [3, 16].

3 Quality of Experience for Multimodal Systems

Although some applications may have optimized QoS parameters, users might still not be completely satisfied. For instance, users may still not be happy with an application as a result of rendering difficulties, entertainment limitations, or cybersickness. In fact, experimental results [64, 65] have shown that systems excelling in QoS do not necessarily translate to an enhancement of the perceived quality due to the gap between system and human-centric evaluations. According to [66], that gap is due to the fact that perception was traditionally treated as multiple modules acting independently. Therefore, the demand to find new methods for a collaborative perceptual quality assessment has increased accordingly [67]. A new model referred to as Quality of Experience (QoE) has been proposed to tackle the issues not revealed by QoS measurements. QoE is defined as:

> A multilevel paradigm of users perceptions and behaviors, representing users emotional, cognitive, and behavioral reactions that are both subjective and objective, while dealing with a multimedia application [67, 68].

Figure 1 depicts the bidirectional correlation between QoE and QoS. QoE depends on both technical influences, that is, QoS, and user influences in terms of cognitive perceptions and behavioral consequences [68]. A study conducted by [67] categorized the user influences into four baskets: perception measures, such as user's satisfaction; rendering quality, for example, cross-modality; psychological measures such as degree of immersion, and physiological measures such as body response. Unlike QoS, QoE evaluation for multimodal system is still immature and little work has been done in this domain [69]. Nevertheless, we itemize three ways to evaluate QoE. The first one is based on mathematical derivations [47, 70]. In this approach, QoE augments QoS but does not totally replace it. Such approach suffers from feasibility and accuracy issues as there is no comprehensive model that can quantify the multidimensionality and large individual variability. The second

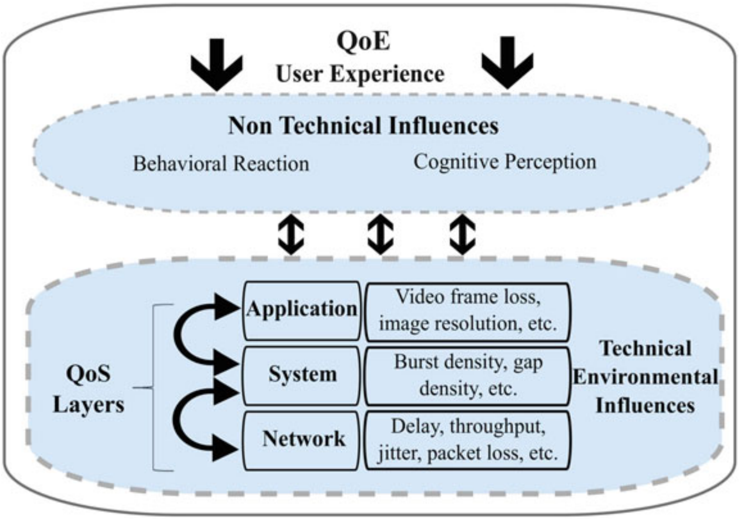


Fig. 1 QoE and QoS dependencies

approach is based on subjective questioners such as the Mean Opinion Score (MOS) [71], in which users explicitly give their opinion about the multimodal system they used. Then, the results are passed through regression analysis to come up with the optimized technical factors that enhance the overall multimedia content. This approach is very expensive, time-consuming, and lacks repeatability. Also, it cannot be applied in real time. The third type produces higher fidelity results and is based on machine learning algorithms, for example, Fuzzy Logic interfaces, that can intelligently anticipate the complicated and vast users' behavior and mentality [67, 70].

Several works that studied QoE in interactive multimodal applications are summarized below. [6, 7] conducted a study based on an online questionnaire to evaluate the effects of user perception of specific multiple-sensorial media (mulsemedia) components like haptic and airflow. Fifty-four users participated in their experiments and filled up their feedbacks accordingly; 70% of the users found that haptic and airflow effects dramatically enhanced their sense of reality and enjoyment level. Whereas, 4% of users experienced some distraction and annoyance, the rest of users gave neutral feedback. This infers the importance of user experience on multimodal applications. However, the experiment was done locally; therefore, the results cannot be extended to tele-multimodal systems where some challenges have to be overcome when disseminating multimodal content over the Internet. The authors addressed this issue in [72]. They use the MPEG-7 description scheme to integrate multiple sensorial metadata to the audiovisual stream. While evaluating this system, a non-precise synchronization between modalities (especially when including olfaction) was reported as the main reason to reduce the user enjoyment levels. This finding was also obtained by [20]. Also, [72] found that reducing some sensorial effects while presenting content to users does not have a negative impact on user perception of the multimedia component. Further, in evaluating what aspects of the mulsemedia content would affect user perception of the delivered content, the results were tactile (62.5%), wind effect (31.25%), with olfaction (6.25%).

The paper [3] examines the cross-modal effect between visual, gustatory, and olfactory via a pseudo-gustatory display outlined above. The authors evaluated the efficiency of their work subjectively by asking participants to try different flavors using Meta cookies, that is, a cookie with neutral taste. Forty-four users experienced six cookie visual/scent combinations. Each user was firstly asked to try plain cookie enhanced with visual and olfactory effects and the second trial without those effects. For more than 79% of the trials, the participants indicate a change in the cookie's taste.

Authors in [70] proposed the IQX Hypothesis Model to outperform the logarithmic function described in [73] that links between the physical stimuli and human perception, in which a natural and generic exponential function (1,2) is used to quantify the relationship between QoS technical parameters (packet loss, jitter, response, and download times) and the QoE factors in terms of MoS findings.

$$\frac{\partial \text{ QoE}}{\partial \text{ QoS}} \sim -(\text{QoE} - \gamma) \tag{1}$$

$$\text{QoE} = \alpha \cdot e^{-\beta \cdot \text{QoS}} + \gamma \tag{2}$$

where:

$$\alpha, \beta, \gamma > 0.$$

The IQX Hypothesis Model has produced a 0.993 correlation coefficient that indicates a desirable matching between the ground truth data and the applied exponential model. This model was only tested on two different types of applications (VoIP and web browsing); therefore, we do not know its effectiveness on environments that have complex combinations of modalities. Further, the model was only evaluated using the (NIST Net) simulator, which has some difficulties in capturing the real IP network behavior [74].

The authors in [47] built a system-level mathematical model for Haptic Audio Visual Environment (HAVE) applications' QoE based on weighted linear combinations of QoS and user experience parameters as depicted in (3).

$$\text{QoE} = \zeta \cdot \text{QoS} + (1 - \zeta\ ) \cdot U\ X \tag{3}$$

where

$$\text{QoE} = \frac{\sum i\ \eta i\ Si}{\sum i\ \eta i} \tag{4}$$

and:

$$UX = A\frac{\sum_i\ \alpha_i\ P_i}{\sum_i\ \alpha_i} + B\ \frac{\sum_j \beta_j\ R_j}{\sum_j \beta_j} + C\ \frac{\sum_k \gamma_k U_k}{\sum_k \gamma_k} \tag{5}$$

where αi, $\beta\ j$, γk and (A, B, C) are the model weighing factors used to maintain the overall quality of experience between 0 and 1. *Si* represents the QoS parameters in terms of delay, jitter, and packet loss, whereas P_i, R_j, and U_k denote the user experience parameters in terms of perception measures, rendering quality measures, and user state measures. Lastly, ζ is used to control the relative priority of the QoS parameters versus user experience parameters. The authors' model was evaluated empirically using subjective testbeds on 30 participants who used a HAVE game called the Balance Ball game [75]. Further, they implemented a Fuzzy Logic Inference System (FIS) using the Mamdani MATLAB that can anticipate the users' QoE based on certain input parameters. Their FIS system maintains a percent error of 4.6% and a correlation coefficient of 0.92.

In [76], the authors discuss a new assessment of olfaction QoE using a novel neuronal model called Functional Connectivity Map (FCM), in which volunteering smellers from Ecole Polytechnique Fdrale de Lausanne (EPFL) were asked to experience different scents while their brain activities were monitored using an electroencephalography (EEG) system. The recorded signals are then passed to the FCM for analysis and further interpretations. The resulting output was used by a machine learning algorithm known as Support Vector Machine (SVM) to classify and anticipate the perceived level of pleasant or unpleasant odor by the participants. As compared to the questionnaires' ground truth, the EEG approach can predict Quality of Experience with up to 65% accuracy rate.

Finally, a recent study [30] tried to define a generic utility model to assess the QoE of multimedia content enhanced with sensory effects. The utility model is also referred to as Quality of Sensory Experience (QuaSE) and is given by (6):

$$\text{QuaSE} = \text{QoE}_{\text{av}}\ \left(\delta + \sum \omega_i\ b_i\ \right) \tag{6}$$

where *wi* indicates the relative weight given to sensory effect, *bi* is binary variable used to confirm whether a given sensory effect has been used or not, and δ represents a fine-tuning parameter. As noticed, the authors claim that the quality of sensory experience (QuaSE) is based on a linear relationship between the QoE of the audiovisual content (QoEav) and the sensory effects number. As far as we know, the utility model is neither implemented nor validated.

In the next section, we identify and classify research methods and techniques for disseminating multimodal information streams while assuring their QoE and QoS requirements.

4 Communication Methods for Multimodal Systems

The networked multimodal environment demands the implementation of a novel, collaborative, adaptive, and interactive communication protocol. To do so, a number of challenges need to be studied and overcome. These challenges can also be described as conditions that the protocol needs to satisfy in order to provide the quality of experience and service required by the end users, modality, as well as any intermediary entities that contribute to the performance of said protocol. The challenges are briefly grouped into *multiplexing, synchronization, reliability, network overhead, interoperability, and adaptability and scalability*. As depicted in Fig. 2, this section highlights these challenges and provides a direction toward the successful implementation of a multimodal communication system.

- **Multimodal fusion** [28, 77, 78] takes into consideration the integration of multiple modalities to be processed. This challenge poses an interesting dilemma that is related to when this fusion of modalities is performed in the overall

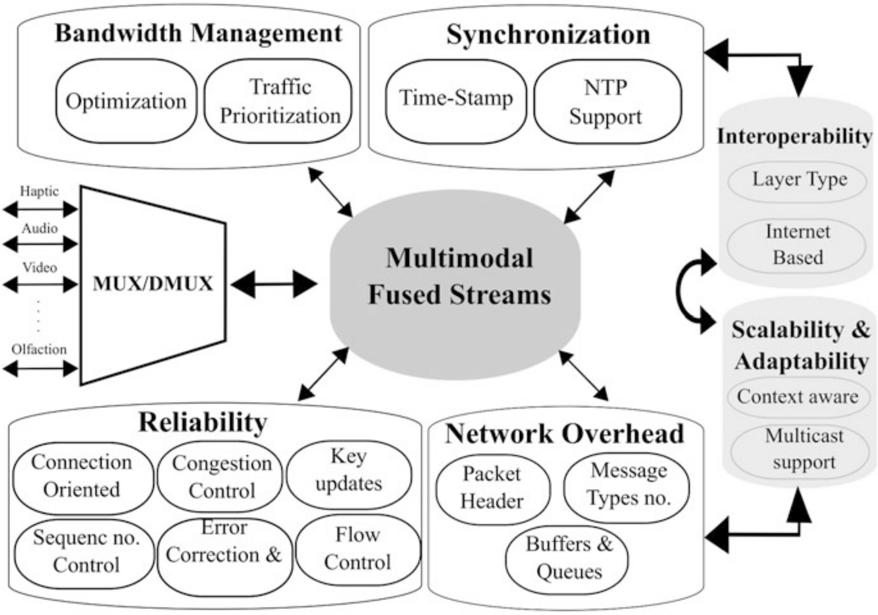


Fig. 2 Higher-level architecture of multimodal protocol

process. Two flavors of this integration have been defined [77]: early and late integration. The former, which also is called feature-level fusion, indicates that the fusion of the data be performed immediately in the beginning of the process, whereas the latter, which is called decision-level fusion, suggests the independent processing of each type of data unimodally prior to being integrated with other information. A hybrid version of these two flavors of multimodal fusion has been introduced as well, referred to as a mid-level integration, which combines some processing and classification techniques before the fusion of data across the different modalities. In this context, the main question that needs to be addressed is to find an effective approach to combine spatial, temporal, proximal, and abstracted unimodal channels of data into one integrated event model.

- **Bandwidth estimation and multiplexing** [79–81] techniques are significant for the protocol implementation in order to better utilize network resources. Bandwidth estimation will allow the protocol to adjust transfer rates based on current network utilization. If the protocol detects that the network is being overloaded, it should lower the transfer rate in order to avoid congestion problems. On the other hand, if the network is being underutilized, the transfer rate needs to be increased to make full use of available resources and provide a better and more efficient use of the bandwidth. Statistical multiplexing, however, makes use of historical statistical information in order to better transmit simultaneous streams of different types of data (haptic, voice, graphics, etc.) which is very important for the multimodal communication protocol. This can be achieved by studying and incorporating upper and lower latency bounds for each modality, as well as a human cross-modal temporal integration, a high degree of impressiveness and transparency. Statistical multiplexing allows for the allocation of bandwidth on a frame-to-frame basis, where bandwidth is supplied to channels according to QoS demands. This technique has proven to achieve better bandwidth utilization [8], and includes approaches such as neural networks [82] and Round Robin [83].
- **Prioritization** [2, 84, 85] is one of the main challenges that has always been the key component to provide end users with a specific quality of experience. Providing end users with seamless and continuous media through a communication protocol can prove to be problematic if streams are not prioritized according to importance and transport protocol capabilities. Determining the appropriate important weight for different streams needs to be taken into consideration in order to make efficient use of the communication medium, as well as provide users with the quality of perception that needs to accompany those streams. For example, real-time applications require a higher priority than video/audio applications to provide the user with the required quality of experience based on perception. Other streams may require different types of data (e.g., haptic, video, and voice) to be transmitted simultaneously, where a prioritization scheme will be required to make sure that data that are, for example, less tolerant of delay and jitter (haptic data for instance) will receive a higher priority than data that have a lower sensitivity to those properties. One way to achieve this is to make use of the IPv4 Type of service (ToS) octet or IPv6 traffic class octet in a Differential Services (DiffServ) architecture in order to provide information about the streams

type/class of service while traversing the communication protocol or it can be implemented by using the description file at the application layer, for example, MPEG7.

- **Synchronization** [11, 44, 86] is another important challenge to consider in order to address different media streams that will traverse the multimodal communication protocol. In fact, some studies have proven that the olfaction stream is very fragile to this condition [20]. In order to achieve an acceptable QoE, video, voice, olfaction, gustatory, and haptic data streams will need to be synchronized accordingly without affecting the end users' perception. Different synchronization techniques have been proposed over the years, which include synchronization buffer schemes, time-stamping and time adjustment algorithms, each with their own uses and requirements. To maintain a synchronized stream of media, three types of levels need to be addressed: (1) intra-stream, (2) inter-stream, and (3) group synchronization control (also known as inter-destination) synchronization. One of the most prominent synchronization schemes, known as adaptive synchronization, can be used to address multiple channels of communication, where both intra-stream and inter-stream synchronization can be taken care of. In addition, according to the findings in [87], adaptive synchronization is immune to clock offset and/or clock frequency drift, does not require a global clock, and can support optimal delay and buffering for given quality of service requirements.
- **Reliability** [2, 8, 88] is another challenge that incurs overhead through interactive protocols that require specific updates of key processes. In order to maintain a certain level of quality of experience, the communication protocol needs to prioritize which updates need to be sent reliably and which updates can tolerate best-effort delivery. In addition, the use of differential messages can be used to minimize the overhead, by sending only pieces of information that were not previously delivered and the receiver cannot estimate. The challenge related to key updates highlights yet another property that the multimodal communication protocol needs to take into consideration, which is the need for minimum overhead during its operation. Some types of data require high overhead due to their nature (e.g., haptics), whereas others incur a lower overhead. Higher overheads can negatively affect the protocol efficiency and need to be addressed accordingly without adversely affecting the perceptual quality of experience. For example, as discussed later, haptics have a refresh rate of 1 kHz, which requires 1000 updates per second. If each update has a header of 20 bytes, this could result in repeated unnecessary overhead close to two megabytes. This could considerably lower the QoE of the end user, and mechanisms need to be implemented into the protocol that will take care of these types of scenarios. Error correction and resilience [11, 89] is a challenge of providing a reliable protocol to end users. Different types of data will require various approaches to provide a satisfactory service to end users, where some can tolerate loss of data and erroneous packets, whereas others require error-free mechanisms. For the latter, different acknowledgment mechanisms have been proposed (e.g., Negative Acknowledgment (NACK) and Selective Acknowledgment (SACK)),

which can reduce the possibility of acknowledgment implosion problems [90] for example. The main problem with supporting error correction and data integrity mechanisms is the extra overhead and latency that can be incurred. Therefore, a trade-off threshold between data integrity and minimum overhead should be considered according to the nature of data, keeping in mind the perceptual quality of experience that end users require based on the application. Minimizing jitter and congestion [41, 88] is also essential for a satisfactory user's QoE. Many solutions have been proposed to tackle this challenge; [91] presented a message multicasting mechanism that uses synchronized clocks for jitter smoothing. This mechanism proved to be easy to implement and independent of the application that uses it, but came at a cost of increased overall delay. Many algorithms have been proposed to be used to reduce congestion, which include additive increase/multiplicative decrease (AIMD) [92], Rate-Based Congestion Control [93], and TCP-Friendly Rate Control [94]. In order to implement a multimodal communication protocol that is efficient, it is important to take these different solutions into perspective, in order to provide end users with a truly real-time interactive quality of experience. Another proposed solution is to optimize the receiver buffer [95] to reduce the unwanted effects of jitter and out-of-order arrival of packets. By introducing an intermediate buffer, the receiving end can consume packets at the right time regardless of time of origin, but can suffer from increased mean delay. This drawback can be taken care of by introducing appropriate receiver buffer length thresholds to accommodate for the potential increase in delay that may occur. Another important control that needs to be implemented is real-time flow control, which is important in the communication protocol as end users could be interacting in a real-time fashion. IP packets traversing the network could be lost, duplicated, or delivered out of order, and this could prove to be problematic to the quality of experience. Sequencing, time stamping, and clock resolution are some of the available proposed mechanisms to be used to achieve real-time flow control.

- **Interoperability** [8, 96] implies that the Internet-based platform is necessary in order to allow for multiple end-to-end systems to interact using this multimodal communication protocol that provides elastic, universal connectivity and standards-based capabilities. This will ease the use of an already existing infrastructure that connects users all over the world together while keeping connection costs at a minimal compared to having a dedicated network set up for all users.
- **Scalability and adaptability** [97, 98] are two major features for the multimodal communication protocol that have to tackle different network conditions. As multimodal communication protocol must connect a huge number of users having multiple modalities into an interactive and collaborative platform, scalability is an important aspect to keep in consideration when building a tolerable communication protocol. In order to provide for this scalability, the protocol needs to adapt to the constantly varying network conditions by monitoring congestion in order to lower/increase transfer rates accordingly. Therefore, multicast communication [99, 100] is indeed an important property that the

multimodal communication protocol needs to provide. This will considerably lower the amount of retransmissions of the same information, utilize the network in a more efficient manner, and mitigate network congestion. IP multicast routing techniques and multi-homing are available and should be incorporated accordingly into the protocol. Adaptability method is to take into consideration the application events and context, which allows for the events in the application to reallocate and adjust resources to different media data accordingly. One interesting finding is that scalability and adaptability are tightly coupled, which means once a scalability is well considered in a protocol design, it will impact positively on the protocol adaptability [101–103].

5 Communication Protocols for Multimodal Systems

As discussed in the previous section, meeting the widely varying communication requirements of each modality, avoiding intra–inter asynchrony of the media streams, steering the network conditions based on the context of the multimedia application, and accounting for heterogeneous nature of each modality are briefly the challenges that need to be tackled in the realization of a novel multimodal communication system. To the best of the authors' knowledge, there are few protocols/frameworks that take into consideration the streaming of the five senses media types while satisfying the aforementioned communication challenges. In Table 2, a number of networking protocols and frameworks are evaluated based on their suitability to handle multimodal communication.

The generic transport-layer protocols, namely, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP), are used by several Internet-based distributed applications. TCP provides reliable and connection-oriented services by employing error, sequence number, loss, and duplication controls. These mechanisms can adversely impact the QoE of any multimedia application. Further, TCP works by creating a virtual connection between two endpoints. Thus, it does not support multicast distribution [104, 105], which means it is not suitable for collaborative class of applications. With a 20-byte header size, TCP creates an extreme network overhead especially while dealing with a modality that requires a high update rate, for example, haptic applications. In general, TCP behavior does not suit any multimedia application since it was created to guarantee a successful delivery of packets, regardless of transmission time. The other widely used generic transport protocol is UDP. It is used by the class of applications that prefers best-effort delivery rather than a reliable data transmission. With a very small header size, 8 bytes, and no retransmission mechanism, UDP has the smallest network overhead among all other protocols. Even so, UDP is not considered suitable for a multimodal class of applications, as described hereafter. First, UDP does not have buffering mechanism, which leads to delay variation. Second, it does not provide an appropriate timing mechanism while sending the datagrams; consequently, the

Table 2 Communication protocols summary

Challenges		Protocols										
		TCP	UDP	LTCP	PR-SCTP	S-SCTP	IRTP	ETP	ALPHAN	Admux	STRON	ADAMS
Scalability and adaptability	Context aware	–	–	–	Yes	Yes	–	–	Yes	Yes	–	Yes
	Multicast support	No	No	Yes	P	Yes	No	–	No	No	–	–
Network overhead	Packet Header (Byte)	20	8	–	12	–	9	–	16[a]	13[a]	–	16[a]
	No. of Messages	1	1	3	2	2	2	1	3	3	2	2
	Buffers	Yes	No	Yes	No	Yes	Yes	–	Yes	Yes	Yes	No
Multiplexing	Bandwidth Optimization	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
	Prioritization	–	–	Yes	Yes	–	–	–	P	Yes	Yes	Yes
	Mixing	No	No	No	Yes	–	Yes	–	No	No	No	Yes
Synchronization	Time-stamp	Yes	–	Yes	–	Yes	–	–	Yes	Yes	Yes	Yes
	Adaptive clocking	No	No	No	No	Yes	No	–	Yes	Yes	–	No
Reliability	Connection Oriented	Yes	No	Yes	Yes	–	Yes	–	No	P	Yes	–
	Congestion Control	CW	No	CW	SAK	NAK	CW	RB	–	P	CW	No
	Flow control	Yes	No	P	Yes	P	P	No	P	P	Yes	No
	Sequence no.	Yes	–	Yes	Yes	Yes	Yes	–	Yes	No	Yes	Yes
	Key updates	No	No	Yes	No	Yes	–	No	–	Yes	–	No
Interoperability	TCP/IP Layer	4	4	4	4	4	4	4	5	5	3	5
	Internet-Based	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes

CW Congestion window, *AK* Acknowledgment, *NAK* Negative AK, *P* Partial, *RB* Rate based, *SAK* Selective AK, **–** Not specified

[a] Layer 4 header not considered

synchronization challenge is not satisfied. Third, UDP is not equipped with a multicast feature, which is needed to achieve the collaboration requirement.

Another protocol referred to as Streaming Control Transmission Protocol [106] is developed to address the limitations of TCP and UDP. In fact, it combines the best features of both TCP and UDP. It resembles TCP in its support for connection-oriented communication; however, SCTP extends the connection concept by establishing multiple transport streams between endpoints within an SCTP port. This process is called SCTP association and is based on the Stream Sequence Number (SSN) in the SCTP header. SCTP has a 12-byte common header that leads each chunk. In addition to unreliable communication, SCTP provides reliable services via the selective ACK (S-ACK) mechanism. Also it can support multi-homing, where the SCTP association can use a pair of IP addresses for each end user. This is very beneficial for load balancing between different paths. SCTP was developed for telephone signaling, however it has been evolved to another transport layer protocol called Partial Reliable-Stream Control Transmission Protocol (PR-SCTP) [107]. The PR-SCTP shares the same size of SCTP header. In fact, it is unreliable data mode extension of SCTP. In this protocol, The data retransmission is varied according to the reliability level of the content within a given association. More specifically, it has several policies that permit to set different reliability weights of data so lost data will be resent until a certain reliability threshold (message lifetime) is reached. When the reliability threshold is reached, the sender ends the unacknowledged data retransmission and notifies the receiver with Forward TSNs in order to ignore any outstanding packets and change the cumulative ACK point toward the news packets. By doing so, the end user can dynamically assign the reliability level of the stream based on the modality type, its requirements, and network conditions. The authors in [108] conducted a performance comparison of PR-SCTP, TCP, and UDP for MPEG-4 multimedia traffic in mobile network. In that study, PR-SCTP gives promising results in terms of maintaining a low delay and bandwidth. Unfortunately, PR-SCTP has not been evaluated for new multimedia systems that include recent modalities such as haptic or 3D graphics.

Light TCP (LTCP) [105, 109] was proposed to overcome the problem of the generic TCP's lack of ability to differentiate between active and outdated messages. Endorsing such a mechanism in the communication protocol will dramatically decrease the amount of updated traffic. This technique is referred to as message obsolescence. Thus, while sending, LTCP provides an updatable queue mechanism to discard the updated message with outdated flag. More specifically, from the sender side a key update message is placed at the end of the queue and a marker is placed beyond it showing its location and indicating its significance compared to regular messages. Obsolete normal message update will be discarded and replaced by new normal messages. This replacement takes place when no update message is required between the older message location and the end of the queue. The messages located in the queue will be aggregated and then sent as a one IP packet according to the congestion control. At the receiver side, the received messages will be passed to the application iff they are newer than the messages that already have been received. As such, LTCP does not provide buffering at the receiver side. LTCP resembles the generic TCP in some aspects such as when the update messages are queued in the sender side, the acknowledgment is per packet sent, both sliding window and congestion control are used, and sequence number is used. LTCP offers key update message and aggregation features compared to the two generic TCP and UDP. For this feature, however, LTCP lacks fast processing, which is highly needed for multimodality interaction communications.

Interactive Real-Time Protocol (IRTP) is developed by [110]. IRTP is a transport protocol that can readapt itself quickly to cater for both crucial and real-time request. For that purpose, it imitates the two generic transport TCP and UDP; TCP is used to transport crucial data (data that has to be delivered to the receiver even if it will incur an extra time delay), whereas UDP is used to transport the real-time stream (data that has to be delivered as fast as possible). For bandwidth probing IRTP adapted two algorithms of flow control, the Sender-Based and Receiver-Based window algorithms. Error control is implemented using the Arrangement Buffer (ARB) and Application Buffer (APB), respectively. It is worth noting that IRTP header consists of four segments: COMMAND, SOURCE IDENTITY, SEQUENCE NUMBER OR ACKNOWLEDGE NUMBER, and CHECKSUM, which result in a header size of 9 bytes. Being a short-headed protocol, it suits the streaming of Internet Robot Control applications where short date per frame is sent frequently. However, [111] have shown that in some congested scenarios, IRTP can create an undesirable delay in bilateral teleoperation.

The Quality-Oriented Adaptation Scheme (QOAS) [29] is an application layer protocol that has been proposed to provide an adaptive client–server multimedia streaming approach over the IP architecture. The protocol addressed the trade-off between number of end users that can simultaneously use the multimedia system and the transmission-related parameters, for example, throughput, packet loss, delay, and jitter, required to deliver a perceived media content. For that purpose, the QOAS sever has four quality versions of the same multimedia content. At the end-user side, a subjective mechanism called Quality of Delivery Grading Scheme (QoDGS) is used to evaluate the quality of the delivered content and submit a feedback report to the server. The server then utilizes a Server Arbitration Scheme (SAS) to filter the received feedback based on the end-user preference and either refines or degrades the quality of the delivered content. The main weakness of the proposed model is that the QOAS content can only be formulated in an MPEG-2 stream, which weakens its usability toward new emerging modalities like haptic and olfaction. Further, the user feedback packet (4-byte payload) is encapsulated into 20-byte-IP header, 8-byte-UDP header, 8-byte-RTCP receiver report packet header that turns to 40-byte overhead for each feedback packet. This scenario might create a bottleneck for collaborative and interactive media applications.

An application protocol, named Application Layer Protocol for Haptic (ALPHAN) [85], uses a mechanism of multiple buffers to prioritize and optimize haptic, VR, and audiovisual transmission. Based on its architecture, ALPHAN is built on top of UDP and uses Haptic Application Meta Language (HAML) as it is a rich meta language that allows customizing of the application requirements. ALPHAN is inspired by the MPEG standard as it classifies the data flow into three main frame types I, P, and B. It has a retransmission method similar to TCP that keeps key updates until they are acknowledged. More specifically, it applies a reliable sending scheme for key updates (I packets), whereas regular updates (P,B) are unaffected. It maintains inter- and intra-stream synchronization by adding the time stamp and sequence number fields to its packet header. Also it uses Multiple Buffering (MB) scheme where each application object is allocating a buffer in the sending side. This allocation provides a distinguish transmission for each object update whether for user-based prioritization or preference-based application. Given the fact that ALPHAN is a layer five protocol, it induces an extra 16-byte overhead for each overpassed UDP packet. This is a drawback especially when no adaptive compression algorithm is used to mitigate the packet rate produced by the ALPHAN application. Lastly, it is not a collaborative enabled protocol since it does not address multicasting. For haptic, visual, auditory, and scent data dissemination, an application protocol called Adaptive Multiplexing Framework for Multimedia communications (Admux) [8] is based on applications requirements and network conditions that define the QoS requirements for haptic, audio, and video. Admux is aimed to use multimedia channels that are fused into a single transport stream. Similar to ALPHAN, Admux uses the HAML so that the multimodal application may tune the Admux communication methods depending on the application's simultaneous requirements. It also uses the MB scheme. Instead of parallel communication, Admux outperforms ALPHAN by using statistical multiplexing that fairly allocates the network resources. The protocol was simulated on an interpersonal telepresence system known as HugMe system and the results showed its adaptability to communication conditions and application events. In general, Admux's header has no multicast field, which implies that it is not suitable for multiuser interactions. Also, the framework induces additional delays due to multilevel fragmentations and packetizations. Furthermore, the authors claim that Admux supports olfactory data; however, the HugMe system simulation was verified on haptic audio video streaming. Also, presuming that the multimodal application has an HAML description file can be a huge weakness that impairs its usability on non-enabled HAML applications and platforms. The last limitation is that its implementation does not show an error resilience algorithm, which is a critical factor for the stability and transparency of any multimodal system.

A protocol called Supermedia Transport for Teleoperations over Overlay Networks (STRON) [112] is designed to provide a forward error correction as a transport scheme that helps in delivering a fast and reliable service. It was created to provide multiple decoupling paths of overlay networks to deliver the packets. Also, STRON outperforms TCP in providing reliable and fast interaction features. By utilizing Reed Solomon codes, STRON offers a transport service that does involve ACK or retransmission in its traffic control. TCP-Friendly Rate Control (TFRC) [94] is a protocol used in STRON to provide a congestion control for each overlay path. Authors in [101] designed a QoS management framework for supermedia teleoperation systems that are sensitive to delay. Using adequate codecs, delay-sensitive streams are encoded and transported through several overlay links. To control transmission rates, the framework uses a transport method that adaptively aggregates haptic packets and applies a priority filtering mechanism. It aims to adapt the haptic transmission, loss rate, and buffering time to network status variations according to the effects of haptic data loss and delay. A synchronization method called deadreckoning is used to offset jitter to a minimum delay. Compared to other transport schemes, the framework has distinguished features, which are priority-based filtering and network-adaptive aggregation of haptic event that are experimentally proven to optimize transmission rates.

For exclusive interactive applications like haptic, there is a protocol named Efficient Transport Protocol (ETP) [113]. By using a feature called Inter-Packet GAP (IGP), which is referred to as the period elapsed between two consecutive packets sent, ETP aims to minimize the round-trip time (RTT) according to the network congestion conditions. By doing so, the instability bandwidth times are detected and then accordingly it will enhance the available bandwidth during transmission. For efficient bandwidth, ETP uses six states of transmission: Fast-Decreasing, Look, Increase IPG, Slow Decrease IPG, Stability IPG, and Stability Max. Utilizing IPG feature leads to providing congestion control in ETP. It separates between flows of the data transport and the feedback channel by using UDP protocol. ETP is suitable for interactive applications where frequently exchanged data like haptic takes place. Although ETP was optimized for such, it does not support important features like multicast, flow control, and key updates, which is its main weakness.

In [72], the authors propose a framework called Adaptive Mulsemedia Delivery Solution (ADAMS) for delivery of video and sensorial data. Their proposal is aimed at streaming a three-dimensional model that includes source of video, sensorial data, and optimization of network bandwidth. Their framework assumes combining three sensorial sources (air motion, haptic, and olfaction) by using MPEG-7 for the description scheme. Although they considered in the framework multiple sensorial sources, their framework is user-dependent, which means the subjects should inform the system administrator of their preferred sensorial effects, in other words it does not address all the possible multimodality combinations. The main weakness in this solution is that it adds an extra 16-byte overhead for each overpassed UDP and TCP packets. This can be a vulnerable bottleneck problem especially when disseminating modalities that need high refresh rate over non-dedicated networks such as the Internet.

6 Concluding Remarks and Prospective Research Avenues

Currently, the Internet-of-Things has been introduced as an umbrella to cover the extensions of the Internet into the physical world by means of the widespread deployment of spatially distributed devices with embedded identification and sensing capabilities. The next wave of innovation demands enabling haptic interaction with audiovisual feedback, as well as technical systems, supporting not just visual interaction, but also that involving robotic systems to be steered and controlled with an imperceptible time-lag. This next wave of innovation will create the Tactile Internet (TI).

Tactile Internet is referred to the Internet works that use multisensory information for the purpose of facilitating multimodal interactions with things (real or virtual) in perceived real time.

In this chapter, we introduce a new class of TI applications, where the ubiquitously multimedia devices are grouped smartly and interconnected with the Internet and other communication networks.

The pervasive interconnection of deployed sensory media devices with existing communication networks, and eventually the Tactile Internet, is referred to as Internet of Multimodal Things.

This includes devices responsible for sensory effects such as olfaction, haptic, and gustatory, and are all combined in a spatial, temporal, and contextual manner. Figure 3 shows a new set of things that include but are not limited to the haptic jacket [2], olfactory display [19, 114], tactile-enabled gloves [115], drones [116], smart armband [117], head-mounted display [118], and biological [119] and physiological sensors [120]. If these sensory devices (i.e., things) are connected cooperatively to the Internet using the new Tactile Internet 5G infrastructure, humans will be able to communicate with other entities (humans, things) interactively using the five common senses and enjoying the ultra-low-end latency. As such, the immersive multimedia experiences might be enhanced/enriched and, consequently, users will have better enjoyment, life, health, and overall well-being.

What are the technologies used to deploy this promising paradigm, how do we implement, deliver, and consume perceived experiences of such systems, and how to combine spatial, temporal, proximal, and abstracted unimodal streams generated from the smart media things into an integrated-event output are beyond the scope of this chapter.

In summary, we provide a comprehensive insight into a class of multimedia systems that are capable of stimulating the five human senses so that high-quality interaction can be achieved. We also explore the benefits of appending haptic, olfaction, and gustatory modalities into the existing audiovisual multimedia systems. We show that the QoS requirements differ based on the modality type and nature. We comment on the up-to-date studies related to QoE frameworks and measurements in the area of interactive multimodal applications. We outline the main research methods required to disseminate multimodal streams while taking into consideration the discussed quality metrics. Furthermore, we show the state of the art on the

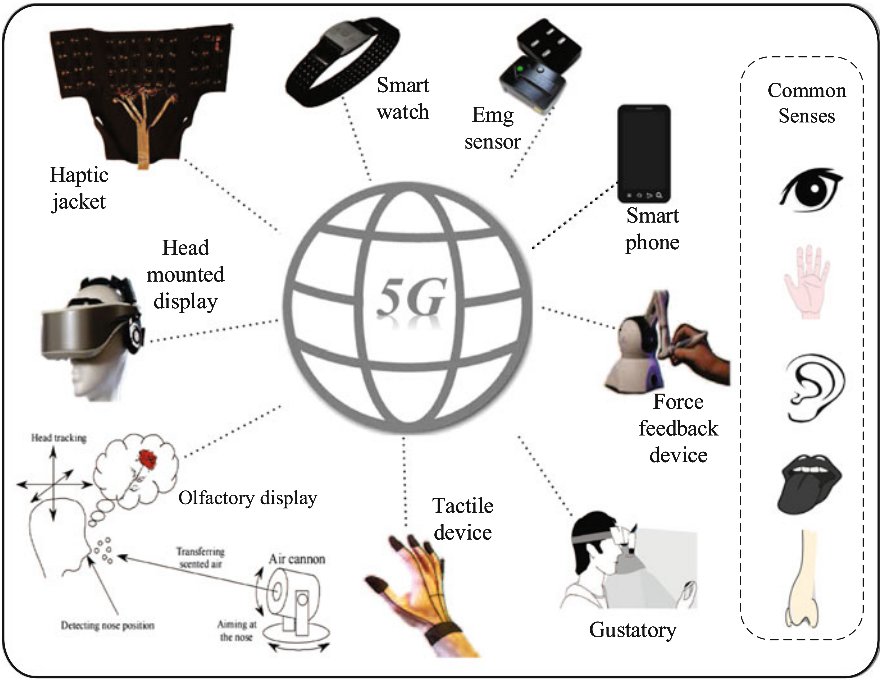


Fig. 3 Internet of Multimodal Things over Tactile Internet

existing networking protocols-frameworks, and we argue their suitability from the multimodal communication perspective. Finally, we complement our discussion on the vision of the Tactile Internet by introducing the Internet of Multimodal Things (IoMT) to enhance the overall quality of life. Although there are several studies that have been conducted on integrating multimodal sensory methods, some areas of research need to be explored in future studies. We highlighted some of the open research avenues as follows:

1. The work toward integrating multimodalities into one system that precisely reflects real-life experience is progressing very slowly, especially when looking deeply on the complex relationship between modalities and their conflicting communication requirements.
2. We show that the integration of different channels of media modalities will involve different parameters of QoS requirements. This implies that the adaptation function to the QoS demands will be based on the number of multiple modalities fusion and their type.
3. We show that HAML and MPEG7 are used to enhance the communication between the multimedia application and the networking techniques. Implementing a new metadata language to describe the temporal, spatial, and content relationship of all modalities is another possible avenue in this research.

4. It has been shown that a new robust transmission paradigm needs to be implemented to send multimodal metadata over strict communications network conditions. Synchronization, multiplexing, and multicasting seem to be the common research themes in this domain.
5. Being the most recent modalities, olfaction and gustatory hardware technologies are still in the immature stages. Thus, more significant research and industrial efforts are needed in this area.
6. More efforts have to be carried out in optimizing a QoE model, for example, automated models, to capture the heterogeneous nature of multimodal systems in a smart and cost-effective way that satisfies the most important entity of the multimedia system (i.e., the user).

The results of this research are expected to be useful for researchers, engineers, and industry working in the area of the Tactile Internet technology.

References

1. Ache, B.W., Young, J.M.: Olfaction: diverse species, conserved principles. Neuron **48**(3) (2005) 417–430
2. El Saddik, A., Orozco, M., Eid, M., Cha, J.: Haptics Technologies: Bringing Touch to Multimedia. Springer Series on Touch and Haptic Systems. Springer Berlin Heidelberg (2011)
3. Narumi, T., Kajinami, T., Nishizaka, S., Tanikawa, T., Hirose, M.: Pseudogustatory display system based on cross-modal integration of vision, olfaction and gustation. In: Virtual Reality Conference (VR), 2011 IEEE, IEEE (2011) 127–130
4. Maier, M., Chowdhury, M., Rimal, B.P., Van, D.P.: The tactile internet: vision, recent progress, and open challenges. IEEE Communications Magazine **54**(5) (2016) 138–145
5. Fettweis, G.P.: The tactile internet: applications and challenges. IEEE Vehicular Technology Magazine **9**(1) (2014) 64–70
6. Yuan, Z., Ghinea, G., Muntean, G.M.: Quality of experience study for multiple sensorial media delivery. In: Wireless Communications and Mobile Computing Conference (IWCMC), 2014 International, IEEE (2014) 1142–1146
7. Yuan, Z., Chen, S., Ghinea, G., Muntean, G.M.: User quality of experience of mulsemedia applications. ACM Trans. Multimedia Comput. Commun. Appl. **11**(1s) (October 2014) 15:1–15:19
8. Eid, M., El Saddik, A.: Admux communication protrocol for real-time multimodal intreaction. In: Proceedings of the 2012 IEEE/ACM 16th International Symposium on Distributed Simulation and Real Time Applications. DS-RT'12, Washington, DC, USA, IEEE Computer Society (2012) 118–123
9. Robles-De-La-Torre, G.: The importance of the sense of touch in virtual and real environments. IEEE Multimedia **13**(3) (2006) 24–30
10. Reed, K.B., Peshkin, M., Hartmann, M.J., Patton, J., Vishton, P.M., Grabowecky, M.: Haptic cooperation between people, and between people and machines. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on, IEEE (2006) 2109–2114
11. Steinbach, E., Hirche, S., Ernst, M., Brandi, F., Chaudhari, R., Kammerl, J., Vittorias, I.: Haptic communications. Proceedings of the IEEE **100**(4) (2012) 937–956
12. Hannaford, B., Ryu, J.H.: Time-domain passivity control of haptic interfaces. Robotics and Automation, IEEE Transactions on **18**(1) (2002) 1–10

13. Salisbury, K., Conti, F., Barbagli, F.: Haptic rendering: introductory concepts. Computer Graphics and Applications, IEEE **24**(2) (2004) 24–32
14. OMNI, P.: Phantom omni haptic device (2016) Online accessed; Monday, February 1, 2017.
15. Kaye, J.J.: Making scents: Aromatic output for hci. interactions **11**(1) (January 2004) 48–61
16. Toko, K.: Biochemical Sensors: Mimicking Gustatory and Olfactory Senses. Biochemical Sensors: Mimicking Gustatory and Olfactory Senses. Pan Stanford (2013)
17. Gardner, J., Bartlett, P.: Sensors and Sensory Systems for an Electronic Nose. Nato Science Series E:. Springer Netherlands (2013)
18. Heymann, E.W.: The neglected sense-olfaction in primate behavior, ecology, and evolution. American Journal of Primatology **68**(6) (2006) 519–524
19. Ghinea, G., Ademoye, O.A.: Olfaction-enhanced multimedia: perspectives and challenges. Multimedia Tools and Applications **55**(3) (2011) 601–626
20. Hoshino, S., Ishibashi, Y., Fukushima, N., Sugawara, S.: Qoe assessment in olfactory and haptic media transmission: Influence of inter-stream synchronization error. In: Communications Quality and Reliability (CQR), 2011 IEEE International Workshop Technical Committee on, IEEE (2011) 1–6
21. Sithu, M., Ishibashi, Y., Huang, P., Fukushima, N.: Ikebana competition in networked virtual environment with haptic and olfactory senses. (Dec 2014) 1–3
22. ISO, .: International standard 5492. sensory analysis vocabulary. ref. no. iso 5492:2008 (e). Technical report (2008)
23. Obrist, M., Comber, R., Subramanian, S., Piqueras-Fiszman, B., Velasco, C., Spence, C.: Temporal, affective, and embodied characteristics of taste experiences: A framework for design. In: Proceedings of the 32Nd Annual ACM NY, USA, ACM (2014) 2853–2862
24. Ferrell, W.R., Sheridan, T.B.: Supervisory control of remote manipulation. Spectrum, IEEE **4**(10) (1967) 81–88
25. Srinivasan, M.A., Basdogan, C.: Haptics in virtual environments: Taxonomy, research status, and challenges. Computers & Graphics **21**(4) (1997) 393–404
26. Helbig, H.B., Ernst, M.O.: Optimal integration of shape information from vision and touch. Experimental Brain Research **179**(4) (2007) 595–606
27. Bresciani, J.P., Dammeier, F., Ernst, M.O.: Vision and touch are automatically integrated for the perception of sequences of events. Journal of Vision **6**(5) (2006) 2
28. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6) (2010) 345–379
29. van Hoven, B.: Multi-sensory tourism in the great bear rainforest. Landabrefid (2011) 19
30. Ghinea, G., Timmerer, C., Lin, W., Gulliver, S.R.: Mulsemedia: State of the art, perspectives, and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **11**(1s) (October 2014) 17:1–17:23
31. Gumtau, S., Newland, P., Creed, C.: Mediate–a responsive environment designed for children with autism. Accessible Design in a Digital World (2005)
32. Systems, M.: Multisensory of olfaciton, visual imagery, vibration and 3d sound (2016)
33. Murer, M., Aslan, I., Tscheligi, M.: Lollio: Exploring taste as playful modality. In: Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction. TEI ’13, New York, NY, USA, ACM (2013) 299–302
34. Lee, J.y., Payandeh, S., Trajkovic, L.: Performance evaluation of transport protocols for internet-based teleoperation systems. Proceedings of OPNETWORK. Washington DC: OPNET Technologies Inc (2010) 1–6
35. Cacheda, R.A., Garcia, D.C., Cuevas, A., Castano, F.J.G., Sanchez, J.H., Koltsidas, G., Mancuso, V., Novella, J.I.M., Oh, S., Panto, A.: Qos requirements for multimedia services. In: Resource Management in Satellite Networks. Springer (2007) 67–94
36. Grot, B., Keckler, S.W., Mutlu, O.: Preemptive virtual clock: A flexible, efficient, and cost-effective qos scheme for networks-on-chip. In: Proceedings of the 42Nd Annual IEEE/ACM International Symposium on Microarchitecture. MICRO 42, New York, NY, USA, ACM (2009) 268–279

37. Wijesekera, D., Srivastava, J.: Quality of service (qos) metrics for continuous media. Multimedia Tools and Applications **3**(2) (1996) 127–166
38. Venkatasubramanian, N., Nahrstedt, K.: An integrated metric for video qos. In: Proceedings of the Fifth ACM International Conference on Multimedia. MULTIMEDIA '97, New York, NY, USA, ACM (1997) 371–380
39. Sabata, B., Chatterjee, S., Davis, M., Sydir, J.J., Lawrence, T.F.: Taxonomy for qos specifications. In: Object-Oriented Real-Time Dependable Systems, 1997. Proceedings., Third International Workshop on, IEEE (1997) 100–107
40. Chen, Y., Farley, T., Ye, N.: Qos requirements of network applications on the internet. Information, Knowledge, Systems Management **4** (2004) 55–76
41. Kokkonis, G., Psannis, K., Roumeliotis, M., Kontogiannis, S., Ishibashi, Y.: Evaluating transport and application layer protocols for haptic applications. In: Proc of. IEEE International Symposium on Haptic Audio-Visual Environments and Games. (2012) 66–71
42. Steinbach, E., Hirche, S., Kammerl, J., Vittorias, I., Chaudhari, R.: Haptic data compression and communication. IEEE Signal Processing Magazine **28**(1) (2011) 87–96
43. Park, K.S., Kenyon, R.V.: Effects of network characteristics on human performance in a collaborative virtual environment. In: Virtual Reality, 1999. Proceedings., IEEE, IEEE (1999) 104–111
44. Hinterseer, P., Steinbach, E., Chaudhuri, S.: Perception-based compression of haptic data streams using kalman filters. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Volume 5., IEEE (2006) V–V
45. Sakr, N., Zhou, J., Georganas, N., Zhao, J., Shen, X.: Prediction-based haptic data reduction and compression in tele-mentoring systems. In: Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE, IEEE (2008) 1828–1832
46. Marshall, A., Yap, K.M., Yu, W.: Providing qos for networked peers in distributed haptic virtual environments. Advances in Multimedia **2008** (2008)
47. Hamam, A., Eid, M., El Saddik, A., Georganas, N.D.: A quality of experience model for haptic user interfaces. In: Proceedings of the 2008 Ambi-Sys workshop on Haptic user interfaces in ambient media systems, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2008) 1
48. Gilbert, E.N.: Capacity of a burst-noise channel. Bell system technical journal **39**(5) (1960) 1253–1265
49. Li, Z., Drew, M., Liu, J.: Fundamentals of Multimedia. Texts in Computer Science. Springer International Publishing (2014)
50. Sayood, K.: Introduction to Data Compression. The Morgan Kaufmann Series in Multimedia Information and Systems. Elsevier Science (2012)
51. Zhu, W., Luo, C., Wang, J., Li, S.: Multimedia cloud computing. Signal Processing Magazine, IEEE **28**(3) (2011) 59–69
52. Li, Y., Sun, J., Shum, H.Y.: Video object cut and paste. ACM Trans. Graph. **24**(3) (July 2005) 595–600
53. Ebrahimi, T.: Quality of multimedia experience: Past, present and future. In: Proceedings of the 17th ACM International Conference on Multimedia. MM'09, New York, NY, USA, ACM (2009) 3–4
54. King, H., Hannaford, B., Kammerl, J., Steinbach, E.: Establishing multimodal telepresence sessions using the session initiation protocol (sip) and advanced haptic codecs. In: Haptics Symposium, 2010 IEEE, IEEE (2010) 321–325
55. Galambos, R., Makeig, S., Talmachoff, P.J.: A 40-hz auditory potential recorded from the human scalp. Proceedings of the National Academy of Sciences **78**(4) (1981) 2643–2647
56. Miras, D., Sadagic, A., Teitelbaum, B., Leigh, J., El Zarki, M., Liu, H.: A survey on network qos needs of advanced internet applications. Internet 2 QoS Working Group (2002)
57. Azuma, R.T., et al.: A survey of augmented reality. Presence **6**(4) (1997) 355–385
58. Hu, S.Y., Jiang, J.R., Chen, B.Y.: Peer-to-peer 3d streaming. Internet Computing, IEEE **14**(2) (2010) 54–61

59. Aljaafreh, M.: An efficient hybrid objects selection protocol for 3d streaming over mobile devices. Master's thesis, University of Ottawa, Canada (2012)
60. Englert, M., Jung, Y., Klomann, M., Etzold, J., Grimmy, P., Jia, J.: A streaming framework for instant 3d rendering and interaction. In: Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology. VRST'15, New York, NY, USA, ACM (2015) 192–192
61. Gracanin, D., Zhou, Y., DaSilva, L., et al.: Quality of service for networked virtual environments. Communications Magazine, IEEE **42**(4) (2004) 42–48
62. Keller, P.E., Kouzes, R.T., Kangas, L.J., Hashem, S.: Transmission of olfactory information for tele-medicine. Interactive Technology and the New Paradigm for Healthcare **18** (1995) 168–172
63. Davide, F., Holmberg, M., Lundstrom, I.: 12 virtual olfactory interfaces: electronic noses and olfactory displays. Communications Through Virtual Technology: Identity Community and Technology in the Internet Age (Pages. 193–220) (2001)
64. Kilkki, K.: Quality of experience in communications ecosystem. J. UCS **14** (2008) 615–624
65. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User acceptance of computer technology: a comparison of two theoretical models. Management science **35**(8) (1989) 982–1003
66. Calvert, G., Spence, C., Stein, B.: The Handbook of Multisensory Processes. A Bradford book. MIT Press (2004)
67. Moller, S., Raake, A.: Quality of Experience: Advanced Concepts, Applications and Methods. T-Labs Series in Telecommunication Services. Springer International Publishing (2014)
68. Wu, W., Arefin, A., Rivas, R., Nahrstedt, K., Sheppard, R., Yang, Z.: Quality of experience in distributed interactive multimedia environments: Toward a theoretical framework. In: Proceedings of the 17th ACM International Conference on Multimedia. MM '09, New York, NY, USA, ACM (2009) 481–490
69. Laghari, K.U.R., Connelly, K.: Toward total quality of experience: A qoe model in a communication ecosystem. Communications Magazine, IEEE **50**(4) (2012) 58–65
70. Fiedler, M., Hossfeld, T., Tran-Gia, P.: A generic quantitative relationship between quality of experience and quality of service. Network, IEEE **24**(2) (2010) 36–41
71. Union, I.: Itu-t recommendation p. 800.1: Mean opinion score (mos) terminology. International Telecommunication Union, Tech. Rep (2006)
72. Yuan, Z., Ghinea, G., Muntean, G.M.: Beyond multimedia adaptation: Quality of experience-aware multi-sensorial media delivery. Multimedia, IEEE Transactions on **17**(1) (2015) 104–117
73. Reichl, P., Egger, S., Schatz, R., D'Alconzo, A.: The logarithmic nature of qoe and the role of the weber-fechner law in qoe assessment. In: Communications (ICC), 2010 IEEE International Conference on, IEEE (2010) 1–5
74. Floyd, S., Paxson, V.: Difficulties in simulating the internet. IEEE/ACM Trans. Netw. **9**(4) (August 2001) 392–403
75. Al Osman, H., Eid, M., El Saddik, A.: Evaluating alphan: A communication protocol for haptic interaction. In: 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, IEEE (2008) 361–366
76. Xu, H., Pereira, F., Timmerer, C., Ebrahimi, T.: Towards quality of sensory experience in multimedia. In Demassieux, N., Campolargo, M., eds.: Proceedings of 2015 European Conference on Networks and Communications (EUCNC), Brussels, Belgium, IEEE (June 2015) 627–628
77. Turk, M.: Multimodal interaction: A review. Pattern Recognition Letters **36** (2014) 189–195
78. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: A review of the state-of-the-art. Information Fusion **14**(1) (2013) 28–44
79. Yan, X., Sekercioglu, Y.A., Narayanan, S.: A survey of vertical handover decision algorithms in fourth generation heterogeneous wireless networks. Computer Networks **54**(11) (2010) 1848–1863
80. Akyildiz, I.F., Lee, W.Y., Vuran, M.C., Mohanty, S.: Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey. Computer Networks **50**(13) (2006) 2127–2159

81. Martin, J.: Telecommunications and the Computer. Number v. 1 in PrenticeHall series in automatic computation. Prentice-Hall (1976)
82. Zhang, G.P.: Neural networks for classification: a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **30**(4) (2000) 451–462
83. Rasmussen, R.V., Trick, M.A.: Round robin scheduling–a survey. European Journal of Operational Research **188**(3) (2008) 617–636
84. Siller, M., Woods, J.: Improving quality of experience for multimedia services by qos arbitration on qoe framework. In: In in Proc. of the 13th Packed Video Workshop 2003, Citeseer (2003)
85. Al Osman, H., Eid, M., Iglesias, R., El Saddik, A.: Alphan: Application layer protocol for haptic networking. In: Haptic, Audio and Visual Environments and Games, 2007. HAVE 2007. IEEE International Workshop on, IEEE (2007) 96–101
86. Serral-Gracia, R., Cerqueira, E., Curado, M., Yannuzzi, M., Monteiro, E., Masip-Bruin, X.: An overview of quality of experience measurement challenges for video applications in ip networks. In: Wired/Wireless Internet Communications. Springer (2010) 252–263
87. Liu, C., Xie, Y., Lee, M.J., Saadawi, T.N.: Multipoint multimedia teleconference system with adaptive synchronization. Selected Areas in Communications, IEEE Journal on **14**(7) (1996) 1422–1435
88. Akyildiz, I.F., Melodia, T., Chowdhury, K.R.: A survey on wireless multimedia sensor networks. Computer networks **51**(4) (2007) 921–960
89. Mitra, S., Seifert, N., Zhang, M., Shi, Q., Kim, K.S.: Robust system design with built-in soft-error resilience. Computer (2) (2005) 43–52
90. Paul, S., Sabnani, K.K., Lin, J.C.H., Bhattacharyya, S.: Reliable multicast transport protocol (rmtp). Selected Areas in Communications, IEEE Journal on **15**(3) (1997) 407–421
91. Gautier, L., Diot, C., Kurose, J.: End-to-end transmission control mechanisms for multiparty interactive applications on the internet. In: INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. Volume 3., IEEE (1999) 1470–1479
92. Yang, Y.R., Lam, S.S.: General aimd congestion control. In: Network Protocols, 2000. Proceedings. 2000 International Conference on, IEEE (2000) 187–198
93. Ohsaki, H., Murata, M., Suzuki, H., Ikeda, C., Miyahara, H.: Rate-based congestion control for atm networks. SIGCOMM Comput. Commun. Rev. **25**(2) (April 1995) 60–72
94. Handley, M., Floyd, S., Padhye, J., Widmer, J.: Tcp friendly rate control (tfrc): Protocol specification. Technical report (2002)
95. Wongwirat, O., Ohara, S.: Haptic media synchronization for remote surgery through simulation. IEEE MultiMedia (3) (2006) 62–69
96. Dalal, S., Patton, G., Shim, H.: System and method for enabling multimedia conferencing services on a real-time communications platform (January 16 2003) US Patent App. 10/167, 712.
97. Sterbenz, J.P., Hutchison, D., Cetinkaya, E.K., Jabbar, A., Rohrer, J.P., Scholler, M., Smith, P.: Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. Computer Networks **54**(8) (2010) 1245–1265
98. Becker, S., Grunske, L., Mirandola, R., Overhage, S.: Performance prediction of component-based systems. In: Architecting Systems with Trustworthy Components. Springer (2006) 169–192
99. Shirmohammadi, S., Georganas, N.D.: An end-to-end communication architecture for collaborative virtual environments. Computer Networks **35**(2) (2001) 351–367
100. Hosseini, M., Ahmed, D.T., Shirmohammadi, S., Georganas, N.D.: A survey of application-layer multicast protocols. Communications Surveys & Tutorials, IEEE **9**(3) (2007) 58–74
101. Cen, Z., Mutka, M., Liu, Y., Goradia, A., Xi, N.: Qos management of supermedia enhanced teleoperation via overlay networks. In: Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on, IEEE (2005) 1630–1635
102. Rowstron, A., Druschel, P.: Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In: Middleware 2001, Springer (2001) 329–350

103. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. SIGCOMM Comput. Commun. Rev. **31**(4) (August 2001) 149–160
104. Stallings, W.: Data and computer communications. Pearson/Prentice Hall (2014)
105. Dodeller, S.: Transport layer protocols for haptic virtual environments. (2004)
106. Stewart, R.: Stream control transmission protocol. RFC 2960 (September 2000)
107. Stewart, R., Ramalho, M., Xie, Q., Tuexen, M., Conrad, P.: Stream control transmission protocol (sctp) partial reliability extension rfc 3758. (2004)
108. Sanson, H., Neira, A., Loyola, L., Matsumoto, M.: Pr-sctp for real time h. 264/avc video streaming. In: Proceedings of the 12th International Conference on Advanced Communication Technology. Volume 1., IEEE (2010) 59–63
109. Kessler, G.D., Hodges, L.F.: A network communication protocol for distributed virtual environment systems. (1996)
110. Ping, L., Wenjuan, L., Zengqi, S.: Transport layer protocol reconfiguration for network-based robot control system. In: Networking, Sensing and Control, 2005. Proceedings. 2005 IEEE, IEEE (2005) 1049–1053
111. Wirz, R., Marin, R., Ferre, M., Barrio, J., Claver, J.M., Ortego, J.: Bidirectional transport protocol for teleoperated robots. Industrial Electronics, IEEE Transactions on **56**(9) (2009) 3772–3781
112. Cen, Z., Mutka, M.W., Zhu, D., Xi, N.: Supermedia transport for teleoperations over overlay networks. In: NETWORKING 2005. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems. Springer (2005) 1409–1412
113. Wirz, R., Ferre, M., Marin, R., Barrio, J., Claver, J.M., Ortego, J.: Efficient transport protocol for networked haptics applications. In: Haptics: Perception, Devices and Scenarios. Springer (2008) 3–12
114. Yanagida, Y., Noma, H., Tetsutani, N., Tomono, A.: An unencumbering, localized olfactory display. In: CHI '03 Extended Abstracts on Human Factors in Computing Systems. CHI EA '03, New York, NY, USA, ACM (2003) 988–989
115. Southhampton, U.: Haptics/tactile devices (2016) Online accessed; Monday, February 1, 2017.
116. Gharibi, M., Boutaba, R., Waslander, S.L.: Internet of drones. IEEE Access **4** (2016) 1148–1162
117. Sosche: Scosche rhythm smart + armband heart rate monitor (2016) Online accessed; Monday, February 1, 2017.
118. WorldViz: Vizmove projection vr system (2016) Online accessed; Monday, February 1, 2017.
119. Akyildiz, I.F., Jornet, J.M.: The internet of nano-things. Wireless Communications, IEEE **17**(6) (2010) 58–63
120. Delsys: Emg sensors (2016) Online accessed; Monday, February 1, 2017.

Automating the Optimization of Web Interfaces for E-Commerce

Aoun Lutfi and Stefano Fasciani

1 Introduction

Retail e-commerce figures have consistently risen over the past two decades. Sales volume is expected to double from $2 trillion in 2016 to $4 trillion in 2020 [1], growing at a faster pace than the traditional brick-and-mortar retail. Indeed, e-commerce accounted for only 8.7% of total retail sales in 2016, while the forecast predicts a 14.6% share in 2020. Mobile e-commerce (or m-commerce) is also on the rise, and the volume of these transactions has matched those processed via personal computers [2]. At the regional level, e-commerce volumes, platforms, and players still present significant discrepancies [3] due to differences in sociocultural patterns and in availability in supporting infrastructures. However, overall trends are positive for all regions, and forecast still predicts a long period of consistent growth.

There is a wide range of factors influencing consumer behavior [4]. Most of these can be mapped to the virtual world of e-commerce, which present new opportunities and challenges compared to traditional retail. The virtual retail world is more competitive as there are far less physical barriers between consumers and businesses, and there is even more competition. With more players coming in every year, small and big players are finding strategies to improve and preserve the engagement with their customers. Most critical success factors are indeed related to IT (Information Technology) aspects of the e-retail process, whereas price is not the dominant factor anymore [5]. Online retail transactions are performed with general-purpose devices such as personal computers or mobile devices through

A. Lutfi (✉)
University of Wollongong in Dubai, Dubai, UAE
e-mail: al702@uowmail.edu.au

S. Fasciani
FEIS, University of Wollongong in Dubai, Dubai, UAE
e-mail: stefanfasciani@uowdubai.ac.ae

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_13

custom-made applications or through the browser, with an ever-increasing shift into mobile [6]. The UX (User Experience) principles [7] that emerged to broaden the scope of HCI (Human–Computer Interaction) are valid in this field as well. Indeed UX elements are vital for any product or service delivered via web [8], and improving the CX (Customer Experience) is drawing more attention and investments from all major online retail firms [9].

The relationship between the HCI elements and CX is complex and case-specific. However, in e-commerce, there are metrics that we can use to evaluate the effectiveness of a specific design. The conversion rate is defined as the ratio between completed transactions and the total number of transactions, including those that had been abandoned. Studies from key players in online retail have found that User Experience (UX) and Graphical User Interface (GUI) have a major impact on the conversion rate. User-friendly GUIs and simple UXs ensure the users are engaged and attached to the website [10–12], and in the case of e-commerce, this helps in improving sales results.

In this chapter, we focus on the user-interaction component of the CX for e-commerce, and in particular on the factors that can be controlled by online retailers. In particular, the user-interaction depends on the interface subcomponents, which include input and output devices. Input devices are usually platform-dependent and hardly controlled by an online retailer, for example, touch-screen for mobile devices, pointers, and keyboard for personal computers. Instead, individual retailers design the GUI of their portal, which has an impact on the conversion rate. Providing user-friendly systems is one of the key objectives for designers of interactive computer-based systems. As technology advances, novel interaction modalities emerge, providing new engagement opportunities as well as new challenges in the design process. HCI studies have demonstrated that it is possible to influence the behavior of humans using computer systems [13, 14]. HCI principles with respect to GUI are broad and not sufficiently detailed to fully guide interface design process, which includes numerous customization possibilities. Therefore, empirical analysis of user data can provide significant insight into the relationship between minor interface elements [15] and CX. Leaving these to arbitrary choices of the designers can have a negative impact on the conversion rate.

Given a reference GUI, which has been reported to deliver higher conversion rates, the process designing the other GUI providing similar performances is a challenging task that requires human intervention. Indeed, programmers must abstract a set of design rules from the reference, then verify their consistency in proposed design, and eventually amend the GUI. These rules include the absolute and relative position of the visible elements of the GUI, their size and color, their interactivity, and the number of elements in the GUI. Large organizations handling large volumes of e-commerce transactions are prone to optimize their GUI down to the last detail because small gains in conversion rates represent significant increases in the sales figures. However, in these organizations, the process of designing, programming, and validating of a GUI for e-commerce involves several individuals with different expertise. Often the professionals involved in coding the GUI are not familiar with HCI principles and vice versa. Minor GUI modification, such as changing the place of a button, despite being simple from the programming perspective, can be lengthy as they require revision and approval from several experts.

In this chapter, we discuss a computational method that supports the design, revision, and amendment of web e-commerce GUI, streamlining the overall process and minimizing the need for HCI experts. In particular, we compute a visual model of a given interface, and we compare it against a set of design rules derived from a reference model. The comparison rates the GUI against the rules and provides a set of recommendations on how to modify the interface to increase the similarity with the reference model. Here we focus only on the layout of the GUI, analyzing absolute positioning of the key interface elements with respect to the frame, and the relative positioning pairs of elements. The generation of the GUI layout model is unsupervised and based only on image processing techniques. HTML code analysis is deliberately not utilized here as it would provide further insights into the layout structure, which are not visible to users. The model is generated exclusively using the browser's rendering of the HTML code, which is what is visually presented to users. However, when multiple optimal reference models exist, we use the HTML code to classify the website and select the most appropriate set of rules to rate the current GUI layout.

In the rest of this chapter, we present an overview of the visual model of the GUI and the steps to compute it and verify it against a set of predefined rules. To prove this approach, we present a study on e-commerce websites where the placement of the checkout button has a significant impact on the online sale process conversion rate. The system identifies nonoptimal placement, and then recommends an alternative position that is likely to improve the conversion rate. The results are validated using the rules to enhance UX when using the Visa Checkout system,[1] which are defined by the service provider.

2 Automated Modeling and Scoring of E-commerce GUI

The process of changing an existing web GUI to a more appealing one is often time-consuming and cumbersome, because it involves several subjects and it requires several stages, such as assessment of current interface, change proposal, approval, implementation, verification, and deployment. The method we describe addresses this issue by reducing the effort of designers in implementing web layout improvements, allowing them to focus on more challenging and higher-level aspects

[1] https://developer.visa.com/capabilities/visa_checkout/docs#

of the UX process. The automation of this process requires executing the following subtasks without supervision:

- Identification of the type of website to select an appropriate reference model or set of GUI design rules
- Computation of a visual model for the current web GUI layout
- Comparison of the model against the reference one, providing a quantitative evaluation of the current design
- Generating a set of recommendation to improve the GUI based on the design rules

In this chapter, we present an application of this approach focused on the placement of the checkout button, which has a significant impact on the conversion rate. The checkout button's optimal placement follows different design rules based on the type of goods being sold online. The four macro tasks listed above have been partitioned and mapped into the following computational modules:

1. Website capture module for extracting HTML code and a screenshot of the rendering
2. HTML-based website classification
3. GUI elements identification via image segmentation and pattern recognition
4. GUI model generation
5. Evaluation of the current layout

The entry point of the optimization process as is visible in the data flow of Fig. 1 is the URL of a website, and multiple set of design rules which must be provided as well. This represents the input for the website capture, which includes both the HTML code (i.e., the metadata) and the screenshot image. These represent respectively the input of the classifier module and the image processing module. The image processing module produces a list of relevant elements in the image, which are fed to the model generation module. The computed visual model and the website type are sent to both the evaluation module and optimization module, which are also provided with a set of website-type-dependent rules on the optimal layout properties. The result of the system is a quantitative evaluation of the original layout, and a set of recommendations for improvement based on the appropriate set of design rules.

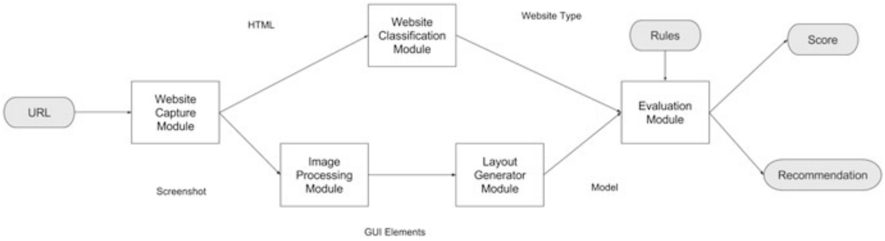


Fig. 1 Flowchart illustrating the data flow and dependencies between the five key modules

2.1 Supporting Techniques

In order to optimize different categories of an e-commerce website, each following a specific set of design rules, it is essential to compare the website layout against the correct reference model. This is automatically achieved using a two-stage classifier that uses keywords from the HTML content to describe the web page. First, the website is scrapped to obtain the keywords, which are then used to perform the classification. For the scrapping, the Document Object Model (DOM) tree-based approach [16] provides satisfactory results, while we select the bag-of-words model for the text-based classification [17, 18]. Since the early 1990s, it has been shown that any approach that involves text classification requires hard categorization of keywords [19]. A bag-of-words approach would satisfy this criterion because it results in a histogram that describes all the words in the text, along with their count.

Also, the identification of interactive elements in the layout, such as buttons and forms, requires two steps. First, the rendering screenshot is segmented and then we use pattern recognition to find which segment is associated with GUI elements. For segmentation, we found that Felzenszwalb's algorithm [20] provides the better performances among those considered in a preliminary study targeting websites images. After segmenting, pattern recognition is used to identify each segment.

Scale Invariant Feature Transform (SIFT) method is a key-point matching algorithm that is scale- and orientation-invariant and capable of matching key-points across multiple images [21]. Although faster pattern recognition algorithms exist, such as Speeded-Up Robust Features (SURF) [22], which is more suitable for real-time applications, SIFT provided higher accuracy when tested, which is the key selection criterion. However, in some cases, key-point matching does not perform well, especially when the template lacks features and complex edges. In such cases, alternative methods used include Optical Character Recognition (OCR) of text (if any text exists) [23] or analysis of color histograms.

For GUI modeling, there are graph-based approaches [24] that can easily be combined with pattern matching. Deep Neural Networks with Markov Random Fields [25] and Long Short-Term memory [26] have been used in image modeling, but they have been shown to work best in regenerative image processing rather than visual-based modeling. Other works in image modeling include the Fisher Vector representation, which uses a set of low-level descriptors to generate a global model [27, 28]. In the literature, there is lack of image modeling based on visual perspective, which can be employed in our system. The importance of the visual-based model is essential in this work. The GUI design highly depends on the human perception of the website layout, which if considered for, would invalidate the model itself and the overall approach.

As mentioned earlier there is a strong correlation between a proper GUI layout and usability [10–14]. Previous works have proposed fuzzy logic to control the GUI and dynamically mutate the layout in real time [29, 30]. These demonstrate that fuzzy logic control is suitable to compare a GUI design, given a set of design rules. The importance of controlling the GUI can also be derived from the three paradigms of HCI [31]. In particular, the first principle states that human–computer interaction is a classical cognitivism and information processing problem that requires the understanding of how one influences the other to properly understand the interaction between both. Understanding this interaction allows some systems to automatically change to adapt, in real time or offline, as we propose in this work.

2.2 Website Capture Module

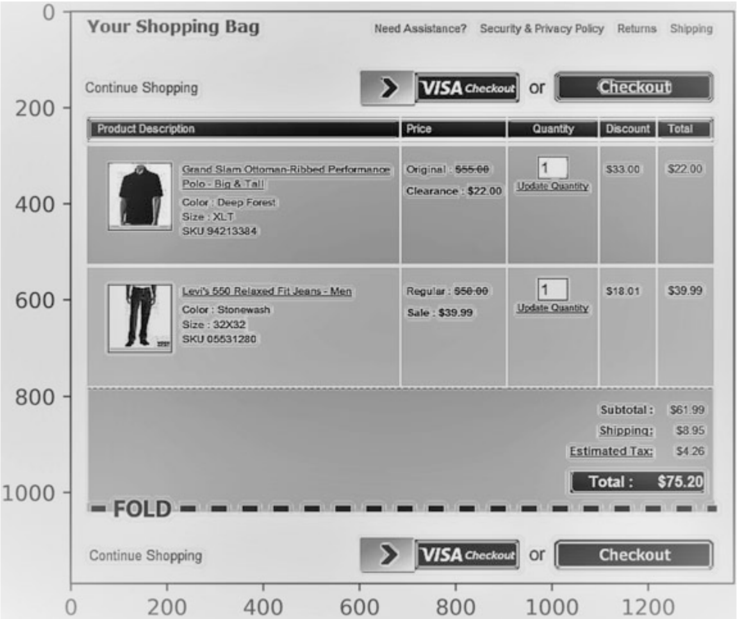
The website capture module takes a website URL, and it fetches the HTML code of the website using an HTTP request. The code is then passed to the website classification module. We use a webdriver (such as *Selenium webdriver*) to render the HTML and obtain an image of the web page. Then the webdriver produces a screenshot of the web page with a parametric geometry and passes it on to the following image processing module. This approach shows its main limitation with websites that dynamically generate some of the elements after delivering the HTML web page. This would result in these elements not being captured in the initial HTML request, which results in a high likelihood of misclassification.

2.3 Website Classification Module

The classifier is based on the bag-of-words approach, which is effective in this context because it provides high accuracy with text-based classification. Before applying the bag-of-words approach, the HTML text is cleaned from tags, scripts, and styles. It is then passed to a natural language processing toolkit to remove common words such as “and,” “or,” “with,” “is.” These words are known as stop-words and are irrelevant for classification purposes. The bag-of-words approach generates a histogram of frequencies for each keyword. Based on this histogram, a score is given to each category based on the number of words assigned to that category. The categories requiring different layout optimization rules in our proof-of-concept application to e-commerce website are shown in Table 1, together with the related keywords used for classification. These categories were also chosen because the great majority of e-commerce websites fall into one of these. As for the keywords, they were selected based on experimental results with a website test set, and are sufficient to describe and correctly classify websites of each category. The classifier then returns the type of website and passes it to the evaluation and optimization modules.

Table 1 E-commerce classification categories and keywords

Category	Keyword
Airlines	Airline, boarding, ticket, tickets, travel, airplane, flight, flights, booking
Hotels	Hotel, hotels, room, rooms, night, nights, booking
Tickets	Ticket, tickets, cinema, show, shows, performance, performances, movie, movies
Food	Food, delivery, meal, meals, order, combo
Generic	Generic, electronics, flowers, fashion, kids, delivery, phone, TV, computer, toy, toys, flower, florist



Your Shopping Bag

Need Assistance? Security & Privacy Policy Returns Shipping

Continue Shopping

Product Description	Price	Quantity	Discount	Total
Grand Slam Ottoman-Ribbed Performance Polo - Big & Tall Color : Deep Forest Size : XLT SKU 94213384	Original : ~~$55.00~~ Clearance : $22.00	1	$33.00	$22.00
Levi's 550 Relaxed Fit Jeans - Men Color : Stonewash Size : 32X32 SKU 05531280	Regular : ~~$50.00~~ Sale : $39.99	1	$18.01	$39.99

Subtotal :	$61.99
Shipping:	$8.95
Estimated Tax:	$4.26
Total :	$75.20

Continue Shopping

Fig. 2 A sample e-commerce website's checkout page segmented using Felzenszwalb's algorithm

2.4 Image Processing Module

This module processes the screenshot of the website's checkout page captured in the previous stage. At first, the algorithm segments the image using Felzenszwalb's algorithm, as in Fig. 2. Then, each segment is analyzed separately to identify whether it contains any GUI-relevant element. Pattern matching is used to identify elements of unique layout, such as the Visa Checkout button that is used for validation. The matching process is based on SIFT to find key-points and on Random Sample Consensus (RANSAC), which uses Homography to find the most likely correct key-points. The module iterates over each segment and tests the

Table 2 Features extracted from the segments key descriptors and the method of extraction

Feature	Method of extraction
Type	From template type
Boundaries	Maximum and minimum key-descriptors positions
Center coordinates	Calculated from boundaries
Dimensions	Calculated from boundaries
Text (if applicable)	OCR
Colors	Color histogram analysis

different templates. To pass the test, templates need at least 10 matched key-points and a statistically sufficient number of well-matched key-points defined by Lowe's ration test [21]. Lowe's test indicates that more than 70% of the key-point descriptor mask must match. Then the segment is localized in the image frame and we extract the features listed in Table 2.

The key-points-based approach works only with elements that contain a larger number of key-descriptors, which are usually detected in edges-rich segments. For simpler elements, such as basic buttons, we take another approach. First, a search for rectangular boxes is performed. Boxes have been shown to represent the majority of buttons and one of the most suitable shapes for a button [10–12]. After discarding all segments that do not contain such elements, the remaining elements are analyzed using OCR, looking for the strings: "Checkout," "Continue," and "Payment." These words usually identify GUI elements for progressing toward the completion of the transaction [10–12]. For product-related images, we take a different approach, performing a color-variance analysis in each segment. Those with high variance likely contain an image of the product being purchased. After performing the analysis on each segment, we extract the features listed in Table 2.

2.5 Model Generation Module

This module computes a graph-based model that describes the relationship between the image elements. The computation is based on the GUI elements extracted at the previous stage. The GUI layout is modeled as a set of nodes connected by links. Each node in the model will describe an element in the GUI. The node structure will be as follows:

- Node Type: (Checkout or Continue or Cancel or Image) (text)
- Node ID: (number)
- Node Coordinates: (x, y) (number, number)
- Node Boundaries: (up, down, left, right) (number, number, number, number)
- Node Dimensions: (h, w) (number, number)
- Node Text: (text)
- Node color: (R, G, B.) (number, number, number)

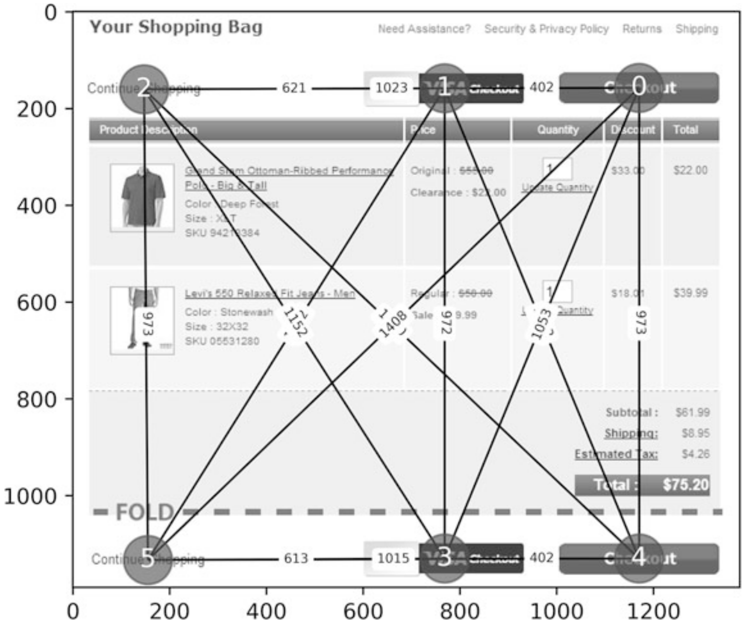


Fig. 3 A sample graphical representation of the model based on the segments in Fig. 2

The connections between each note consist of the distances between each center and the direction, which results in a two-dimensional vector. Any node can have an arbitrary number of links to all the other nodes, describing the relative placement of each. Consequently, comparing any two models will involve comparing the links and nodes. As such, comparing two links can only be done if both links connect two nodes of the same type. The comparator computes differences in the distance and direction (a difference vector). The comparator also calculates the differences between nodes as well; two "checkout" nodes can be compared, and the result would also be a comparison between the different node components (text, dimensions, coordinates, and color). Figure 3 shows the resulting model after the segmentation performed as in Fig. 2.

As is visible in Fig. 3, the links describe the relationship between each relevant element whereas each node describes each relevant element. This modeling technique represents only elements that are relevant and visible to a human observer, which is a significant advantage over those proposed in the literature. However, extracting such information can be challenging: experiments, tests and data analytics could indicate which elements are more relevant to users, depending on the application context. In e-commerce, when aiming at optimizing the conversion rate, it has been observed that the positioning of the checkout button, continue shopping button, back button, and the list of shopping items have the greatest impact on the conversion rate [10–14].

2.6 Evaluation Module

After computing the nodes and links of the model, the evaluation module processes it according to the website type and according to the provided design rules. These rules are parsed according to the following syntax:

1. *Color intensity* of X **with respect** to Y
2. *Location* of X **relative to** Y (above, below, right of, left of)
3. *Text* of X **contains** "text"
4. *Distance* **between** X and Y
5. X *alignment* **with respect to** Y (center align, right align, left align)
6. *Size* of X **with respect to** Y (same, larger, smaller)

X and Y in the rules indicate two different nodes (elements) identified by the node type and an optional identifier, the keywords in **bold** help the parser identify the two nodes in question, and the words in *italic* help identify the property being measured. Both X and Y use the node type as a method of identification since generally in e-commerce websites elements, such as the checkout button, are not redundant in the GUI. The evaluation module utilizes these rules to assess the generated model. This is done comparing each parsed rule with all nodes and links, and checking if these nodes and links match the rule. If so, a score is incremented. Nonetheless, the system also requires a small margin of error to account for minor inaccuracies in the system. In the current implementation, the evaluation is presented on a scale ranging from 0 to 100. The evaluation module also computes the differences between the desired reference model generated by the rules and the actual model and then uses the differences to generate a set of recommendations to improve the GUI. The recommendations indicate which rule was broken and show the required changes to satisfy the rule. If the changes are implemented by the programmer, the GUI is likely to provide higher conversion rates.

3 Implementation and Experimental Results

The method discussed above was implemented as open-source software[2] written in Python. In the partially optimized current implementation, modules are executed as parallel threads when possible, and verbosity level of the screen output can be controlled to reduce the execution time. For the website capture, we use the Python libraries *urllib2* to retrieve the HTML code and *Selenium webdriver* to get a screenshot of the website rendering. However, due to the dynamic scope and different configurations of individual websites, the automatic navigation to

[2] https://github.com/aounlutfi/E-commerce-Opimization

the checkout page was not always possible. Therefore, we include an extra input parameter that represents the URL of the checkout page.

In the classification module, the HTML code obtained is preprocessed using *bs4 BeautifulSoup*, *re*, *collections*, and *nltk*. *Bs4* and *re* are used to clean the code from tags, scripts, and styles. And the natural language processing toolkit *ntlk* is employed to remove stop-words. The histogram is generated using the *collections Counter* method. The implementation of this module has shown 100% accuracy with a test set of 20 valid URLs and three outlier URLs. As mentioned before, the limitation of this approach is the incompatibility with dynamic HTML documents. Any change in the HTML after the code is fetched is not considered for evaluation.

The image processing module is the most computationally intensive module. The average execution time is about 90 seconds across all websites. The module integrates *OpenCV* (for SIFT and Homography), *PIL Image* (for image transformations), and *skimage* (for image segmentation). The OCR functionality is implemented using the *tesseract-OCR* wrapper of the Tesseract OCR engine. Image processing produced an overall accuracy of 95%. This is due to some text elements being identified as buttons. The pattern matching was able to detect all Visa Checkout buttons with 100% accuracy.

The accuracy of the computed model with respect to the website layout is strictly dependent on the performance of the element identification in the image processing module. Table 3 shows the model accuracy for each of the nine samples used to assess the system. As it can be seen in the table, most of the samples were 100% accurate. Three samples present a false positive, mostly due to elements (especially text) visually similar to buttons. These false buttons can be considered constructively toward the GUI improvement goal. In fact, these suggest that the website may contain unclear visual elements that can be mistakenly identified as buttons by some users. The dataset used to assess the image processing and modeling algorithms are composed of nine sample screenshots, one of which is an ideal implementation (template) of the Visa Checkout button.

Table 3 Image processing and modeling results

Sample	Correct number of elements	Obtained number of elements	Model accuracy (%)	False positives
Sample 1	3	3	100	-
Sample 2	4	4	100	-
Sample 3	3	3	100	-
Sample 4	6	4	100	-
Sample 5	2	3	66	1
Sample 6	4	7	57	3
Sample 7	2	2	100	-
Sample 8	1	1	100	-
Sample 9	4	3	50	1

After computing the model of the website's layout, the system evaluates it against the following set of design rules:

- Element 1 is Visa Checkout
- Element 2 is checkout
- Element 1 right of element 2
- Element 1 less than 600px from element 2
- Element 2 contains "checkout"

As expected, the template Visa Checkout implementation scored 100%, whereas other implementations resulted in a lower score. It is interesting to note that although some implementations were correct for their own website type when compared to this specific set of design rules for the "generic" website type, they produced a lower score. This highlights the importance of the classifier to identify the type of website using the correct set of rules for that type. The novelty of this approach and the lack of similar modeling techniques of graphical layout make it difficult to compare our system to others. However, our evaluation of the system is reliable as it was carried out using reference data assessed by experts.

3.1 Further Development

The GUI optimization approach we described in this chapter, including the graph-based visual model and the syntax of the design rules support the simultaneous optimization of multiple interactive elements. However, this determines an increased computational complexity as the number of combinations to verify may increase exponentially. The current implementation supports only the identification via image processing of simple elements such as buttons. The system requires more advanced techniques—to identify complex visual structures such as a table—which are often used in the checkout pages of e-commerce websites. The evaluation module could be extended to return an amended version of the HTML code implementation instead of producing a set of recommended changes. Finally, this approach can be exported to those contexts, beyond web e-commerce, where there have been studies determining measurable aspects of the UX with the layout of the GUI. The method we developed is generic and may find application in a wider spectrum of scenarios. The method can be extended to model and adjust specific features of any graphical interface. The role and nature of the image acquisition, classification based on metadata, image processing, and model generation would not differ as they can be used to describe the relevant elements in other interactive systems.

The method we described can also be used to determine a set of optimal design rules, especially if correlated with a large amount of data companies store for all online transactions [32]. Usually, these datasets are not publicly available as they are considered a valuable asset. However, data analytics can reveal the most successful interfaces deployed over the years in terms of user engagement or conversion rate. These can be modeled with the proposed techniques and from the resulting graph we can produce a set of design rules. Moreover, a deep analysis including all deployed GUIs, including the less successful ones, enable to gain further insights on the relationship between GUI elements and UX. This approach can be applied also to further customize the UX at the single-user level. Indeed, data analytic can already reveal preferences and use patterns within specific categories of users sharing similar profiles. Pushed to an extreme, we can assume to determine and then dynamically apply user-specific HCI design guidelines to improve system usability at the individual level. This aligns with the emerging trend of using machine learning to create the personalized CX. Artificial intelligence is already used in e-commerce for marketing purposes. Indeed recommendation systems based on user profiling information [33] has determined significant sale increases. Implementing a dynamically changing UX can also help in generating the set of standard design rules based on which UI/UX provides a better conversion rate.

4 Summary

In this chapter, we presented a generic method to optimize the design of web interfaces by finding the placement of interactive elements that maximize measurable usability parameters. The method has been implemented as an open-source software and customized to automatically optimize the position of the checkout button in various types of e-commerce websites. The algorithm computing the GUI layout model is based on the combination of several image processing techniques. Results have shown that the system operates with a high degree of accuracy and we discussed a further generalization to widen the application of this technique within and beyond the context of e-commerce. The current method produces a set of recommendations that if included in a revised version of the GUI will determine a better CX and likely higher sales volume. The changes are still to be implemented manually by programmers. However, as discussed in the chapter, the complexity of this process of GUI design and amendment is the identification of changes and approvals. These usually require the intervention of one or more professionals with specific expertise and can be a time-consuming process especially for large organizations. Here we delegate this task to algorithms running on machines, significantly reducing requirements in terms of time and human expertise. Finally, we discussed how to extend this method to work with more complex interfaces, and how to use it to determine the set of design rules to implement a GUI for maximizing a specific usability parameter in the UX.

References

1. eMarketer (2016) Worldwide Retail Ecommerce Sales Will Reach $1.915 Trillion This Year
2. Internet Retailer (2016) 2017 Mobile 500

3. International Post Corporation (2016) State of e-commerce: global outlook 2016-21
4. Solomon MR (2016) Consumer Behavior: Buying, Having, and Being, 12 edition. Pearson, Boston
5. Feindt S, Jeffcoate J, Chappell C (2002) Identifying Success Factors for Rapid Growth in SME E-commerce. Small Business Economics 19:51–62. doi: https://doi.org/10.1023/A:1016165825476
6. Meola A (2016) The Rise of M-Commerce: Mobile Shopping Stats & Trends. In: Business Insider. http://www.businessinsider.com/mobile-commerce-shopping-trends-stats-2016-10. Accessed 13 Dec 2017
7. Hassenzahl M, Tractinsky N (2006) User experience - a research agenda. Behaviour & Information Technology 25:91–97. doi: https://doi.org/10.1080/01449290500330331
8. Garrett JJ (2010) The Elements of User Experience: User-Centered Design for the Web and Beyond. Pearson Education
9. Batra MM (2017) Customer Experience–An Emerging Frontier in Customer Service Excellence. Competition Forum 15:198–207
10. Harshman C (2013) A/B Test Ideas for E-Commerce Call to Action Buttons. In: Optimizely Blog. https://blog.optimizely.com/2013/12/07/ab-test-ideas-call-to-action-buttons/. Accessed 24 Oct 2016
11. Holst C (2011) Fundamental Guidelines Of E-Commerce Checkout Design. Smashing Magazine
12. Laja P (2014) How to Design an Ecommerce Checkout Flow That Converts. In: ConversionXL. http://conversionxl.com/how-to-design-an-ecommerce-checkout-flow-that-converts/. Accessed 24 Oct 2016
13. Gould JD, Lewis C (1985) Designing for usability: key principles and what designers think. Communications of the ACM 28:300–311
14. Shneiderman B, Plaisant C (2004) Designing the User Interface: Strategies for Effective Human Computer Interaction, Fourth. Pearson Addison Wesley
15. Song J, Zahedi F (2001) Web Design in E-Commerce: A Theory and Empirical Analysis. ICIS 2001 Proceedings
16. Kadam VB, Pakle GK (2014) A Survey on HTML Structure Aware and Tree Based Web Data Scraping Technique. International Journal of Computer Science and Information Technologies 5:1655–1658
17. Harris ZS (1954) Distributional structure. Word 10:146–162
18. Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1:43–52
19. Sebastiani F (2001) Machine Learning in Automated Text Categorization. ACM Computing Surveys
20. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. International Journal of Computer Vision 59:167–181
21. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International journal of computer vision 60:91–110
22. Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In: European conference on computer vision. Springer, pp 404–417
23. Smith R (2007) An overview of the Tesseract OCR engine. In: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. IEEE, pp 629–633
24. Foggia P, Percannella G, Vento M (2014) Graph Matching and Learning in Pattern Recognition in the Last 10 Years. International Journal of Pattern Recognition & Artificial Intelligence 28:1. doi: https://doi.org/10.1142/S0218001414500013
25. Wu Z, Lin D, Tang X (2016) Deep Markov Random Field for Image Modeling. In: European Conference on Computer Vision. Springer, pp 295–312
26. Theis L, Bethge M (2015) Generative image modeling using spatial lstms. In: Advances in Neural Information Processing Systems. pp 1927–1935
27. Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: Theory and practice. International journal of computer vision 105:222–245

28. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:13126034
29. Agah A, Tanie K (2000) Intelligent graphical user interface design utilizing multiple fuzzy agents. Interacting with computers 12:529–542
30. Chen S-M, Tan J-M (1994) Handling multicriteria fuzzy decision-making problems based on vague set theory. Fuzzy Sets and Systems 67:163–172. doi: https://doi.org/10.1016/0165-0114(94)90084-1
31. Harrison S, Tatar D, Sengers P (2007) The three paradigms of HCI. In: Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems San Jose, California, USA. pp 1–18
32. Kohavi R, Rothleder NJ, Simoudis E (2002) Emerging trends in business analytics. Communications of the ACM 45:45–48
33. Schafer JB, Konstan JA, Riedl J (2001) E-Commerce Recommendation Applications. In: Applications of Data Mining to Electronic Commerce. Springer, Boston, MA, pp 115–153

Arabic Multimedia Search Platform

Mohamad Raad, Majida Bayan, Yihya Dalloul, Majd Ghareeb, and Amin Haj-Ali

1 Introduction

The work described in this chapter was initiated because of the need to access specific video segments from multiple videos from a single hyperlinked document. That is what the system described herein provides for both English language videos and Arabic language videos.

By "video," the authors mean a multimedia file that has both a moving picture component and an audio component (and not just a moving picture component as many "video" experts would use the term). The system described is also applicable to a purely digital audio file.

There are multiple use cases where such a system could be useful. Two use cases include the use of such a system by journalists to locate relevant archived material and its use by students searching through educational video content. With the introduction of video as a learning aid at multiple levels of schooling [1], and the corresponding increase in the volume of video available, it becomes imperative to create an easy-to-use means for organizing that content into a useful knowledgebase.

The need for fast video search tools can be discerned from the results presented in [2], where a number of everyday computer users were asked to search for parts of a given video and their search strategy recorded. Most adopted a start to finish strategy (i.e. they started at the beginning of the video and browsed it until the end), which obviously is time-consuming and would not be a practical approach for searching a large video database. The discussion presented in [3] provides more insight into the need for tools that allow both directed and undirected search. The latter being the use case where the user is exploring a given database without a specific target in

M. Raad (✉) · M. Bayan · Y. Dalloul · M. Ghareeb · A. Haj-Ali
School of Engineering, Lebanese International University, Beirut, Lebanon
e-mail: mohamad.raad@liu.edu.lb; amin.hajali@liu.edu.lb

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_14

mind. Developing more advanced and interactive video search tools is of significant importance for the extraction of knowledge from video and that is the objective of the work described in this chapter.

2 Related Work

Organizing video for search engines may be achieved in multiple ways. One popular way is the use of metadata. This data is typically based on the information provided by the creator of the video or by user feedback. However, using metadata means being dependent on the content producer to accurately document that content, which is not an ideal situation. For example, a similar system to the one we present has been reported in [4], where a systems integration approach has also been taken (in a similar manner to the system we describe) for the development of an archive search engine (AXES) that provides text, image, video and audio search. The reported precision values are quite low for on-the-fly search as well as for the pretrained-model-based search. However, the AXES system provides for search by example as well. All of which is in reaction to the lack of knowledge depth that metadata can provide. However, as with many similar systems, the AXES system is focused on European languages. Another similar system was presented in [5], where the focus was on the development of a video retrieval system for lecture content. The system applied OCR as well as ASR to identify relevant videos and to search for relevant video segments. The authors reported a positive impact on the learning outcomes for users of the system versus non-users of the system, although they do not report how the abilities of the users for the different systems were normalized. Again, the focus of this system was a European language (German).

An application similar to the focus of our work is described in [6], where a database of mini-lectures is transcribed automatically and the transcription is available for editing by the lecturer to ensure accurate information capture. This is a medium-sized system with thousands of such lectures and where the concern is with the accurate capture of spoken information in text format. In our system, the concern is the identification of relevant videos based on the audio and allowing the user to determine the usefulness of the video, whereas in [6], the main focus is enabling the content author to refine the content to enhance its searchability. Similarly, the authors of [7, 8] developed a similar system to the one described in our work, with the main differences being our use of images to identify relevant people (most likely speakers) in the video stream, as will become clearer in the latter parts of this chapter. Of course, the focus of these systems was also on a European language whereas our focus has been on the Arabic language. A tangentially similar work is an interesting application presented in [9], where the authors had the aim of tracing the speech learning curve of newborn children by determining when specific words had been mentioned in the presence of the child for the first time and how much influence that had on the way the child uses that word. It is interesting to note that the authors reported that at the time the automated speech recognition systems were inadequate for the task and so human transcribers were required, whereas most recent efforts aimed at capturing human speech have become automated with varying degrees of success reported.

The broader scope of the work presented herein is information connectivity. This concept is presented in [10], which discusses the benefits of “associative browsing” to enhance the knowledge extracted from electronic information sources. This is a conceptual extension of information hyperlinking, a subset of which is video hyperlinking. Hyperlinking video is becoming an increasingly important tool for information retrieval; for example, [11] describes the results of Trechvid 2016 which focused on video information retrieval. One of the themes in that workshop was video hyperlinking, whilst another theme was ad hoc video search. The event shows the significance of developing tools for information retrieval from video; however, it seems to have a broader scope than knowledge enhancement and stretches to person and object tracking. An example video hyperlinking system that has been patented and has commercial application is described in [12]. Another approach to hyperlinking videos based on audio content has been described in [13]. The aim of that work was to link similar videos together rather than serve search results, although the similarity between the two objectives is high. The work again focuses on one of the European languages (English).

The use of audio for the linking of multimedia content requires a mapping from one information domain (sound) to another (text) as text remains much easier to handle for search engines or other types of software-based processing. This mapping is typically completed automatically, although the authors of [14] concluded that manual annotation is still the best way of hyperlinking video data with the introduction of other information such as context and audio producing a reduction in performance. Such studies indicate that well-structured multimedia information will be much more easily browsed than badly structured multimedia. This may be more of an indication of the lack of effectiveness of the tools than of the requirement for better multimedia structuring. Similarly, a study of video hyperlinking techniques in [15] seemed to conclude that there is some way to go before these techniques become reliable for knowledge enhancement. In any case, many other researchers have gone down the path of automated mapping using Automatic Speech Recognition (ASR) or transcription.

Transcribing multimedia archives has been an important topic for some time and of course it is not just focused on enabling the search of video content; for example, [16] describes an effort to mine transcribed call center conversations for knowledge to enhance the service offered by the call center. This is an example of how such a system as developed in our work can have a practical knowledge focused application. The error margin of transcription, however, seems to be the major concern for the authors since that can lead to gaps in the knowledge extracted. We disagree with that conclusion. Many books may cover the same topic; one does not need to read all of them to develop a deep understanding of that topic. Similarly, one does not need to identify all the audio segments where a specific topic is mentioned to extract useful knowledge, although the enhancement of the transcription engines would clearly enhance the knowledge extracted from such recordings.

In a similar manner, the work in [17] highlights the difficulties that transcription systems encounter, such as signal noise, the bandwidth used for the recording and the language models used to extract transcription data. Many commercial ASR systems are available but there are many free, open source options as well. The authors of [18] present a comparison of available open source speech recognition and transcription systems, concluding that SPHINX [19] is an easily deployable speech recognition system even if it is not the best system (Kaldi [20] was found to be the best performing system but the slowest to integrate) and so is a suitable choice for a system integration task. A practical example of the difficulties that transcription-based search systems will have in identifying relevant content can be found in [21], where the need for the development of a less subjective language than English for communicating with robots has been addressed. Transcription-based search systems rely on the identification of search terms as entered by the searcher. Each language has a different level of subjectivity in typical use and certainly when applied to highly subjective languages such as English or Arabic, transcription systems need to be enhanced with additional features to be useful. Similarly, [22] reports on the difficulty of obtaining consensus on the best transcription of an audio track amongst human experts, which further highlights the difficulty of identifying a highly accurate speech transcriber for adoption in a transcription-based search or hyperlinking.

A key component of developing ASR is the development of transcribers and properly constituted language models. To produce such models, speech corpora in each target language is required. There are many such useful corpora, for example [23] describes an Arabic speech corpora for multi-dialect Arabic to be used for research in Arabic language focused research. The corpora consist of Gulf, Levantine, Formal (Modern Standard Arabic) and Egyptian Arabic. The corpora are transcribed and are available for use by emailing the author on the address available at [24]. Another Levantine Arabic corpus is available at [25]. There are similar copora for other languages, for example [26] describes a Thai news broadcast corpus, the New Zealand English corpus is described in [27], the casual French corpus in [28], a pan-European news broadcast corpus is described in [29] and a Mandarin news broadcast corpus is available in [30].

A popular tool for the production of speech corpora in multiple languages is “Transcriber” [31, 32], which allows for the manual annotation of speech files and the development of XML documents transcribing the speech information. This tool has been superseded by “TranscriberAG” [33], which has enhanced features for the ease-of-speech corpus production. Such tools are useful for the development of dictionaries that can be used as input to the development of automated transcription tools. A related work is reported in [34] that allows for the reduction of the time needed to manually annotate speech files. Again, these tools are useful in enhancing the quality of transcribed speech. The development of such tools has taken multiple paths; for example, [35] describes a JAVA library that allows the development of applications that use temporal and structural information for the analysis of speech and language instead of providing a stand-alone tool.

These tools are critical to the development of systems that aim to make it easier to identify useful information contained within multimedia files. The work described in the following sections are an extension of the system described in [36], which was an initial attempt by some of the authors at developing a hyperlinked video system. In that case, the video files remained on-line; the SPHINX [19, 37] transcriber was used to generate the time-stamped transcription of the target files followed by the generation of the relevant hyperlinks based on the keyword search. The links addressed segments within the YouTube database and the videos. That system was then extended for use with the Arabic language.

3 System Design and Implementation

The Arabic multimedia search platform was developed to meet the following two main use cases:

1. Transcribing an audio file and generating an hyperlinked index
    - 1.1. The system is given the audio file as input, where the audio is assumed to be in WAV format.
    - 1.2. The system is given a list of phrases to search for.
    - 1.3. The system searches the audio file for each word or phrase and generates a timed index for each identified occurrence of each word or phrase.
    - 1.4. The system creates a hyperlink for each identified occurrence.
    - 1.5. The system generates a hyperlinked file listing all the hyperlinked occurrences of the phrases.
2. Transcribing from a video file and generating a hyperlinked index.
    - 2.1. The system is given the video file as input where the video is assumed to be in MP4 format and the audio (.WAV) is extracted from the video.
    - 2.2. The system is given a list of words or phrases to search for.
    - 2.3. The system is given an image to refine the search results by. The uploaded image is used for an image search within the uploaded video.
    - 2.4. The system searches the audio file for each word or phrase and generates a timed index for each identified occurrence that corresponds with the time frame of the identified location of the image being searched for.
    - 2.5. The system creates a hyperlink for each identified occurrence to the relevant video segment.
    - 2.6. The system generates a hyperlinked file listing all the hyperlinked occurrences of the phrases with respect to filtered images.

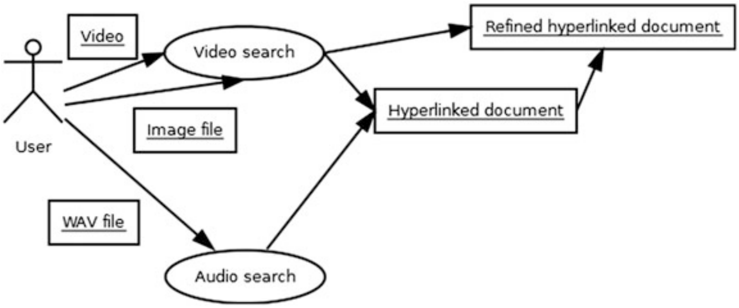


Fig. 1 System use cases

Figure 1 summarizes the main use cases for the developed system. Based on the above use cases, the following requirements were identified for multimedia files:

1. In the case where the audio file is the input:
   (a) The system shall accept a WAV file as input, which is an uncompressed audio format.
   (b) The system shall be able to transcribe speech data.
   (c) The system shall be able to search transcribed data for phrases.
   (d) The system shall accept a list of search words or phrases as input.
   (e) The system shall associate each word or phrase occurrence with a location within the accepted WAV file.
   (f) The system shall hyperlink each phrase occurrence to the original audio file.
   (g) The system shall generate an XML-based file that lists, as hyperlinks, the pointers for each identified occurrence.
2. In the case where the video file is the input:
   (a) The system shall accept an MP4 file as input.
   (b) The system shall extract the WAV file from the MP4 file.
   (c) The system shall be able to search transcribed data for words and phrases.
   (d) The system shall accept a list of phrases as input.
   (e) The system shall associate each phrase with a location within the accepted WAV file.
   (f) The system shall hyperlink each phrase occurrence to the audio file.
   (g) The system shall generate an XML-based file that lists, as hyperlinks, the pointers for each identified occurrence.
3. In the case where the video file and an image file are the inputs:
   (a) The system shall accept an MP4 file as input.
   (b) The system shall accept a JPG file as input.
   (c) The system shall extract the WAV file from the MP4 file.
   (d) The system shall be able to search transcribed data for phrases.
   (e) The system shall accept a list of phrases as input.

(f) The system shall compare uploaded image with database images for image filtering process.
(g) The system shall associate each phrase and image occurrence with a location within the accepted WAV file.
(h) The system shall hyperlink each phrase and corresponding image occurrence to the audio file.
(i) The system shall generate an XML-based file that lists, as hyperlinks, the pointers for each identified occurrence.

To meet the above-listed requirements, the system had to have the components shown in Fig. 2. The main components, as shown, are the web interface, the search engine, the transcriber, video frame extractor, image search engine and hyperlinked document generator. The web interface allows the user to add multimedia files to be transcribed so that the database has tagged transcription results added to it. The hyperlinking engine is driven by the search terms and so as more searches are conducted using the system, more hyperlinked results are maintained in the database. Each hyperlinked document is referred to as an "index" since it links search terms to locations within the body of multimedia information available. The first version of an index may be refined via the image search engine. Basically, the hyperlinks linking to instances of a search term mentioned in the multimedia

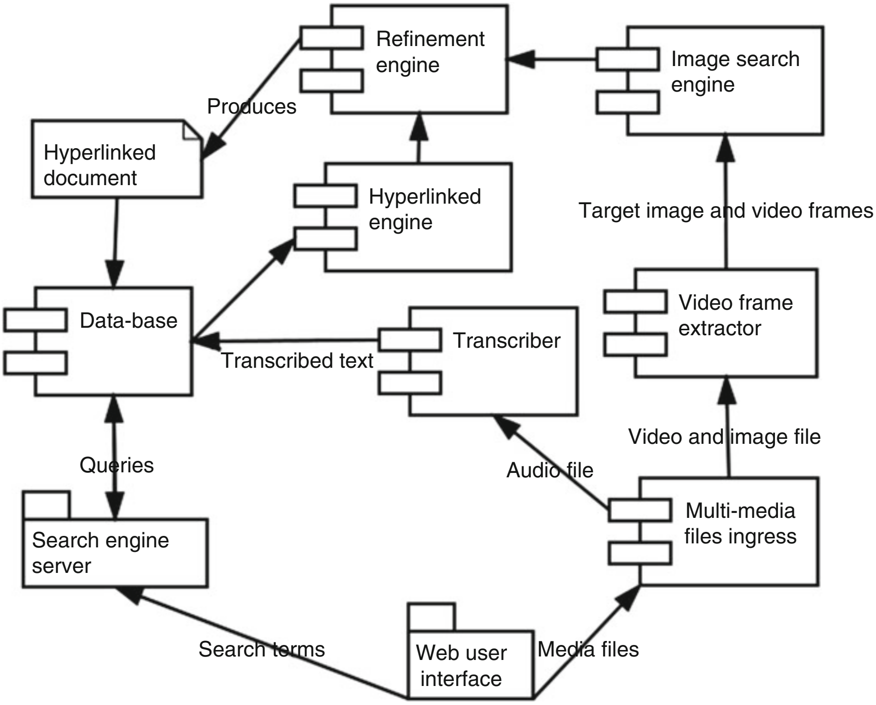


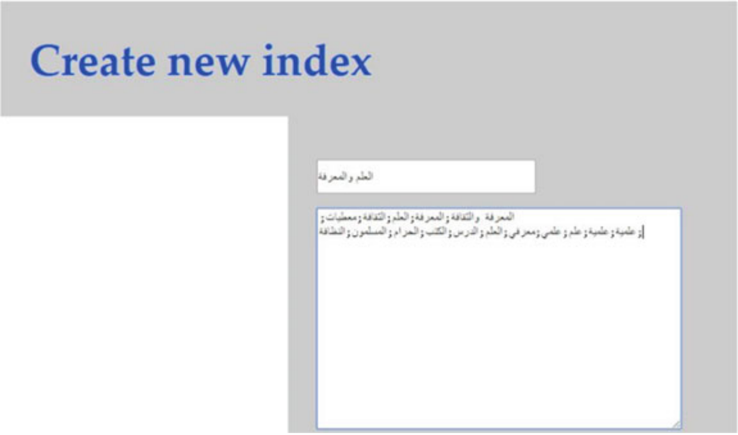
Fig. 2 High-level system architecture

files available are used to determine if the searched for image matches objects in the video frames that are in the temporal vicinity of the search term. If it does, the hyperlinks are added to the second version of the index. Both versions are maintained in the database for future use.

The multimedia ingress component in the current implementation handles both video and audio files. This component is based on FFMPEG [38] because of FFMPEG's wide-ranging functionality, community of users and because it is open source. Clearly, this component is not as efficient as it may be; however, it has been reliable. The focus of our work has been on MP4 files and WAV files. When a video file is uploaded by the user, the ingress component extracts the audio track and sends it to the transcriber. In our system, the transcriber is based on the IBM Watson speech-to-text cloud service [39]. The choice to use a cloud service was made after reviewing a number of the available transcription tools. Informal tests showed that this service was the most accurate for formal (MSA) Arabic as well as English. However, for English-based content, the system also has CMU SPHINX built in for transcription. The transcribed files are maintained in a DB for future use as more and more searches are conducted, and so more hyperlinked documents are generated.

The system uses the OpenIMAJ [40] image search library to identify objects within a video based on the image input by the user. The focus of our experiments has been on face recognition, namely to identify the presence of a person (not necessarily the speaker) in the video at the time a given word or phrase is uttered. To date, all our experiments have been informal but the OpenIMAJ library has been found to be very reliable for this task. The image search is not conducted over the entire set of frames but rather over a 1 s period after the detection of a "searched for" term. In other words, this is localized search.

Figure 3 shows the current interface allowing the user to specify the Arabic search terms. Figure 4 shows the output of the ingress process which has used

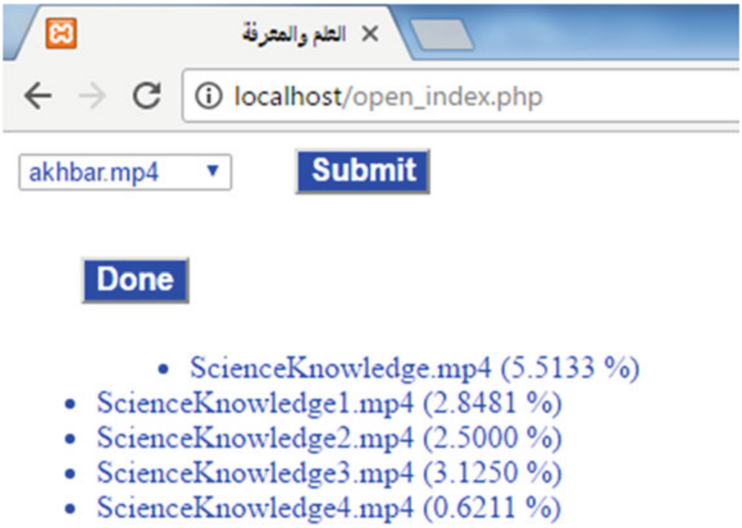


Create new index

العلم والمعرفة

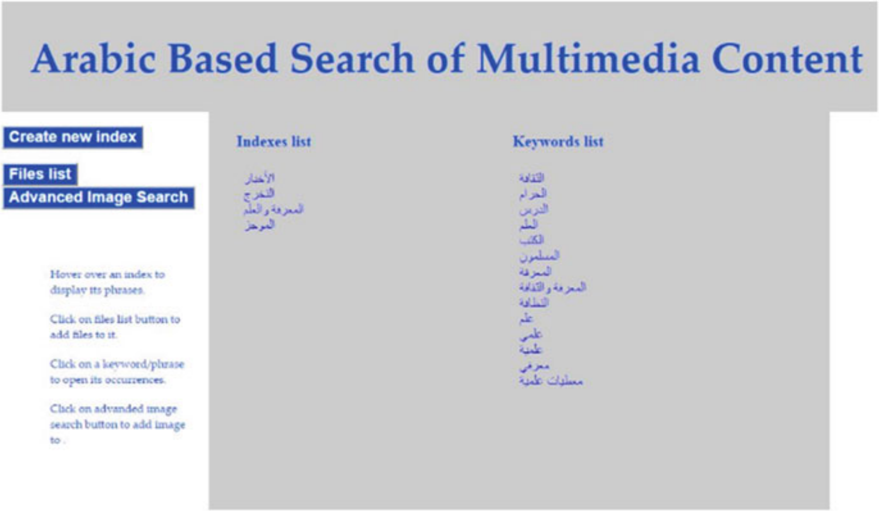
المعرفة والثقافة ;المعرفة;العلم ;الثقافة;معطيات ;
;علمية;علمية ;علم ;علمي;معرفي;العلم;الدرس;الكتب;الحرام;المسلمون;النظافة

Fig. 3 Creating a new hyperlinked document (index)—the search terms entered



- ScienceKnowledge.mp4 (5.5133 %)
- ScienceKnowledge1.mp4 (2.8481 %)
- ScienceKnowledge2.mp4 (2.5000 %)
- ScienceKnowledge3.mp4 (3.1250 %)
- ScienceKnowledge4.mp4 (0.6211 %)

Fig. 4 Adding new videos to the DB



Indexes list

الأخبار
التخرج
المعرفة والعلم
الموجز

Keywords list

الثقافة
الحرام
الدرس
العلم
الكتب
المسلمون
المعرفة
المعرفة والثقافة
الطاقة
علم
علمي
علمية
معرفي
معلومات علمية

Hover over an index to display its phrases.

Click on files list button to add files to it.

Click on a keyword/phrase to open its occurrences.

Click on advanded image search button to add image to .

Fig. 5 Example hyperlinked document

the specified search terms to create an index based on the uploaded files. The percentages shown in brackets next to each video indicate the relevance of the video based on the specified search terms. Figure 5 shows an example hyperlinked document where each term is linked to a series of hyperlinks indicating where that term has occurred in different videos. Figures 6 and 7 show an example refined search result based on the image in Fig. 6.

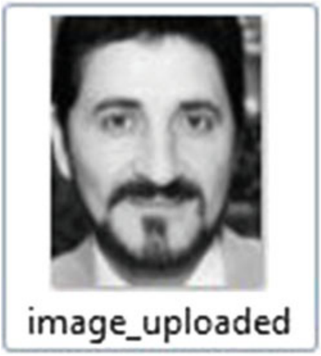


Fig. 6 Example search refined according to this image

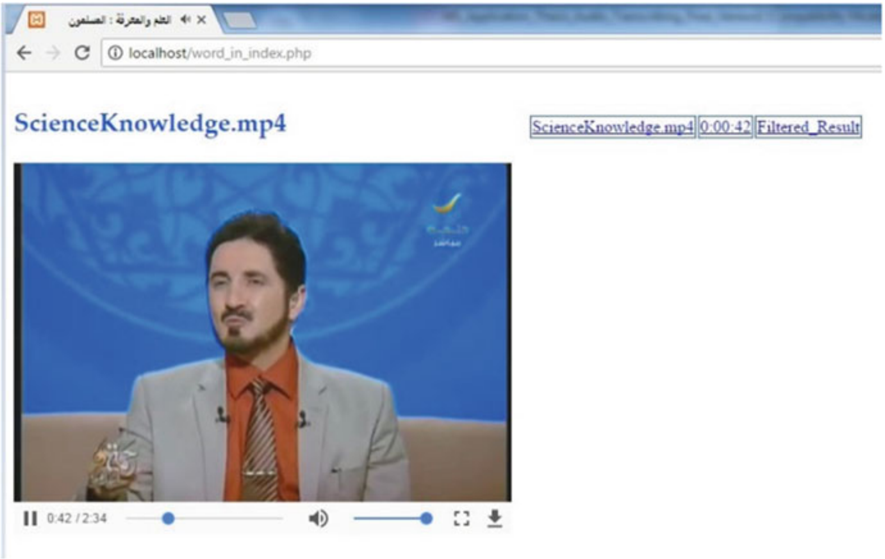


Fig. 7 Example refined result showing playback

4 Discussion and Future Work

The work described in this chapter has been developed using a system of systems approach. The different components of the system are almost all open source and stable releases. The one standout is the Arabic transcription engine, which is a commercial cloud service. The main reason is that the open source or free transcribers were found to underperform this service.

The main issue with the system in its current state is that the Arabic language has multiple dialects and is used in a subjective manner [41, 42], meaning that the models used to transcribe the audio have to be evolved and more rigorously tested. This will be the next phase of developing the system presented in this chapter. Yet another difficulty has been the identification of an objective measure for determining the performance of our system and similar systems. Whilst other works mentioned previously have relied on Word Error Rate (WER) or some other "ground truth"-based measure, this is not very relevant in the authors' opinion to the objective of the system. The system is intended to enhance the ability to access and use relevant multimedia content. The objective measure should reflect that aim instead of simply focusing on whether all utterances of a particular word have been captured.

The main purpose in developing this system was to help extract knowledge from Arabic-based multimedia content. By developing these types of tools, Arabic content would remain accessible and useful and hence the Arabic language would be maintained as a language for knowledge transfer. All the reviewed systems are focused on European languages. Whilst their retasking to another language may seem to only require changing the transcription language, that is actually a major change from the system development perspective, specifically from the accuracy and relevance point of view given the specifics of the Arabic language. Finally, given that the developed system has both an Arabic and English component, another future direction of this work will be to hyperlink content in both languages, allowing for a search in one language to also return results in another language, allowing knowledge extraction from a wider range of sources.

Acknowledgement The work presented in this chapter was kindly supported by a CNRS-L "Grant Research Programme 2016" (Lebanon) fund.

References

1. J. Burgess and J. Green, "Uses of YouTube: Digital literacy and the growth of knowledge," in *YouTube: Online video and participatory culture*UK: Polity press, 2009, pp. 126–143.
2. K. Schoeffmann and C. Cobârzan, "An evaluation of interactive search with modern video players," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, 2013, pp. 1–4: IEEE.
3. K. Schoeffmann and F. Hopfgartner, "Interactive video search," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1321–1322: ACM.
4. T. Tommasi *et al.*, "Beyond metadata: searching your archive based on its audio-visual content," in *IBC2014*, Amsterdam, Netherlands, 2014, pp. 1–3.
5. H. Yang and C. Meinel, "Content based lecture video retrieval using speech and video text information," *IEEE Transactions on Learning Technologies,* vol. 7, no. 2, pp. 142-154, 2014.
6. J. D. V. Miró, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan, "Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories," *Speech Communication,* vol. 74, pp. 65–75, 2015.
7. M. Eskevich *et al.*, "Multimedia information seeking through search and hyperlinking," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 2013, pp. 287–294: ACM.
8. M. Eskevich, R. Aly, D. Racca, R. Ordelman, S. Chen, and G. J. Jones, "The search and hyperlinking task at MediaEval 2014," 2014.

9. R. Kubat, P. DeCamp, B. Roy, and D. Roy, “Totalrecall: visualization and semi-automatic annotation of very large audio-visual corpora,” in *ICMI*, 2007, vol. 7, pp. 208-215.
10. S. Kairam, N. H. Riche, S. Drucker, R. Fernandez, and J. Heer, “Refinery: Visual exploration of large, heterogeneous networks through associative browsing,” in *Computer Graphics Forum*, 2015, vol. 34, no. 3, pp. 301–310: Wiley Online Library.
11. G. Awad *et al.*, “Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking,” in *Proceedings of TRECVID*, 2016, vol. 2016.
12. A. Chakraborty, P. Liu, and L. Hsu, “Method and apparatus for authoring and linking video documents,” USA Patent US6462754 B1, Oct 8, 2002.
13. P. Galuščáková and P. Pecina, “Audio Information for Hyperlinking of TV Content,” in *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia*, 2015, pp. 27-30: ACM.
14. Z. Cheng, X. Li, J. Shen, and A. G. Hauptmann, “Which information sources are more effective and reliable in video search,” in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 1069–1072: ACM.
15. R. J. Ordelman, M. Eskevich, R. Aly, B. Huet, and G. Jones, “Defining and evaluating video hyperlinking for navigating multimedia archives,” in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 727–732: ACM.
16. M. Garnier-Rizet *et al.*, “CallSurf: Automatic Transcription, Indexing and Structuration of Call Center Conversational Speech for Knowledge Extraction and Query by Content,” in *LREC*, 2008.
17. C. Barras, A. Allauzen, L. Lamel, and J.-L. Gauvain, “Transcribing audio-video archives,” in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, vol. 1, pp. I-13–I-16: IEEE.
18. C. Gaida, P. Lange, R. Petrick, P. Proba, A. Malatawy, and D. Suendermann-Oeft, “Comparing open-source speech recognition toolkits,” *Tech. Rep., DHBW Stuttgart,* 2014.
19. P. Lamere *et al.*, “The CMU SPHINX-4 speech recognition system,” in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, 2003, vol. 1, pp. 2–5.
20. D. Povey *et al.*, “The Kaldi Speech Recognition Toolkit,” presented at the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011.
21. O. Mubin, C. Bartneck, L. Feijs, H. Hooft van Huysduynen, J. Hu, and J. Muelver, “Improving speech recognition with the robot interaction language,” *Disruptive Science and Technology,* vol. 1, no. 2, pp. 79–88, 2012.
22. B. C. Roy and D. Roy, “Fast transcription of unstructured audio recordings,” presented at the The 10th Annual Conference of the International Speech Communication Association, INTERSPEECH, Brighton, UK, September, 2009, 2009. Available: http://hdl.handle.net/1721.1/67363
23. K. Almeman, M. Lee, and A. A. Almiman, “Multi dialect Arabic speech parallel corpora,” in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, 2013, pp. 1–6: IEEE.
24. K. Almeman. (December 25). *Arabic resources*. Available: http://www.almeman.com/arabic-resources.html
25. N. Halabi. (2017, December 25). *Arabic speech corpus*. Available: http://en.arabicspeechcorpus.com/
26. A. Chotimongkol, K. Saykhum, P. Chootrakool, N. Thatphithakkul, and C. Wutiwiwatchai, “LOTUS-BN: A Thai broadcast news corpus and its research applications,” in *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*, 2009, pp. 44–50: IEEE.
27. R. Fromont and J. Hay, “ONZE Miner: the development of a browser-based research tool,” *Corpora,* vol. 3, no. 2, pp. 173–193, 2008.
28. F. Torreira, M. Adda-Decker, and M. Ernestus, “The Nijmegen corpus of casual French,” *Speech Communication,* vol. 52, no. 3, pp. 201–212, 2010.

29. A. Vandecatseye *et al.*, “The COST278 Pan-European Broadcast News Database,” in *LREC*, 2004.
30. H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, “MATBN: A Mandarin Chinese broadcast news corpus,” *International Journal of Computational Linguistics and Chinese Language Processing,* vol. 10, no. 2, pp. 219–236, 2005.
31. C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, “Transcriber: development and use of a tool for assisting speech corpora production,” *Speech Communication,* vol. 33, no. 1, pp. 5–22, 2001.
32. K. Boudahmane, M. Manta, F. Antoine, S. Galliano, and C. Barras. (2008, December, 25). *Transcriber*. Available: http://trans.sourceforge.net/en/presentation.php
33. DGA. (2014, December, 25). *TranscriberAG*. Available: http://transag.sourceforge.net/
34. B. Bigi, “SPPAS-multi-lingual approaches to the automatic annotation of speech,” *Phonetician,* vol. 2015-I-II, no. 111-112, pp. 54 - 69, 2015.
35. J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann, “The NITE XML toolkit: flexible annotation for multimodal language data,” *Behavior Research Methods,* vol. 35, no. 3, pp. 353–363, 2003.
36. M. N. Al Laham, I. Ayass, M. Ghareeb, Z. El-Bazzal, and M. Raad, “Audio indexing for YouTube,” in *Digital Information and Communication Technology and its Applications (DICTAP), 2015 Fifth International Conference on*, 2015, pp. 111–114: IEEE.
37. W. Walker *et al.*, “Sphinx-4: A Flexible Open Source Framework for Speech Recognition,” Sun Microsystems, Inc., Technical Report 2004.
38. (2017). *FFMPEG*. Available: http://ffmpeg.org/
39. IBM. (2017). *IBM Watson speech to text cloud service*. Available: https://www.ibm.com/watson/services/speech-to-text/
40. J. S. Hare, S. Samangooei, and D. P. Dupplaw, “OpenIMAJ and ImageTerrier: Java libraries and tools for scalable multimedia analysis and indexing of images,” presented at the Proceedings of the 19th ACM international conference on Multimedia, Scottsdale, Arizona, USA, 2011.
41. N. V. Yushmanov, *The structure of the Arabic language*. 1961.
42. K. Versteegh, *The arabic language*. Edinburgh University Press, 2014.

MOALEM: An Assistive Platform for Children with Difficulties in Reading and Writing Arabic

Jihad Mohamad Alja'am, Moutaz Saleh, Dominic Massaro, and Mohamad Eid

1 Introduction

The number of children with learning difficulties (LD) is increasing in the whole world. A significant number of Qatari schoolchildren are facing difficulties in reading, understanding and writing Arabic words and sentences. Many of them stop learning and become somehow isolated and marginalized. In fact, in 2016, it was estimated that 20% of children in elementary schools in the country experience LD.[1] The current methods of teaching are not suitable for them. In addition, teachers do not have enough resources and time to introduce properly new concepts in the classrooms. Parents cannot afford the high cost of special education centers and private instructors to teach their children. Therefore, teaching these children is recognized as a serious and alarming problem that needs immediate intervention. They need special attention, dedicated care, and tailored curricula that suit each one of them. They should learn to communicate and rely on themselves in reading, understanding, and writing.

[1] Ministry of development planning and statistics, https://www.mdps.gov.qa/en/statistics1/pages/topicslisting.aspx?parent=Social&child=SpecialNeeds 2016.

J. M. Alja'am (✉) · M. Saleh
Computer Science and Engineering Department, Qatar University, Doha, Qatar
e-mail: jaam@qu.edu.qa; moutaz.saleh@qu.edu.qa

D. Massaro
University of California, Santa Cruz, CA, USA
e-mail: massaro@ucsc.edu

M. Eid
New York University—Abu Dhabi, Abu Dhabi, United Arab Emirates
e-mail: mohamad.eid@nyu.edu

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_15

The aim of this research work is to develop and evaluate theoretical models and practical strategies for the use of multimodal multimedia (combining visual, auditory, and haptic modalities) to help children with LD overcome their learning problems. To achieve our goal, we propose an assistive platform entitled MOALEM[2] (Multimedia-Oriented Arabic Language Educational Materials) that uses multimedia technology, text mining, haptic and avatar tools. These software and hardware components have the potential to revolutionize the manner of learning and teaching in schools and special education centers. The proposed algorithms can automatically mine Arabic texts in a specific domain, extract the main concepts (actions, event, actors, etc.), and link them directly to multimedia elements. Children can then see images that explain the appropriate meaning of the Arabic vocabulary occurring in the text. Teachers can address logical questions to MOALEM and get intelligent responses through an inference engine. We use Arabic text processing techniques including morphological analysis and ontology [1]. MOALEM possesses a haptic (touch-based) device to teach children how to write Arabic words correctly by physically guiding their hand along a sequence of strokes of reference handwriting. MOALEM has an avatar called “Badr” that can pronounce standard Arabic words. Children can listen to and watch Badr to observe the correct pronunciation of words and learn the corresponding articulation and emotional expression. The usability of the MOALEM platform is not limited to special education centers and schools. Families can use MOALEM as well to assist their children at home. MOALEM can also be used to teach Arabic to nonnative learners who want to learn Arabic as a second language.

The chapter is organized as follows. Section 2 discusses the background. Sections 3 and 4 present the MOALEM platform. Finally, Sect. 5 concludes the chapter.

2 Background

The population of children with LD is noticeably increasing across the globe [2, 3]. These children require comprehensive care and quick intervention especially during the early childhood years [4]. They struggle with learning new concepts and can sometimes display a negative behavioral attitude. It is possible that they can successfully learn with different methods than those used with their normal peers. Thus, it is very important to understand the difficulties that these children are facing and develop a customized curriculum with personalized contents to accomplish remedial learning and put them on a productive path for new learning.

There are two types of learning disabilities, as detailed in [5]: (1) global learning disability (GLD) and (2) specific learning disability (SLD). Children with GLD have difficulties in appropriately understanding almost everything new as others would. These children are called “slower learners” in the literature and

[2] The acronym MOALEM is an Arabic word (معلم) meaning teacher, educator, or instructor.

their thinking abilities are below average of their normal peers. Children with SLD are of average intelligence and they need different teaching methodologies to help them understand. They can continue as normal students if the appropriate teaching methods are found to suit their effective needs. These children are currently integrated in normal schools in the state of Qatar, and this adds key strains on these schools and its teachers. In fact, these children need more dedication and one-to-one teaching methodology, but schools cannot recruit new teachers to follow children individually.

Generally, children have problems in several language skills that include the following: (1) *reading problems*: for instance, they cannot distinguish properly among letters and words, different words, forming simple proper sentences; (2) *writing problems*: they have difficulties in writing letters and words properly, including writing letters out of order or in reverse; and (3) *understanding problems*: they may not be able to understand the correct meaning of simple words in each context. In addition, they may have difficulties understanding simple sentences correctly, forgetting names of objects and locations, or establishing logical relations between words. These difficulties can be due to different physical and psychological problems, for instance, intellectual disabilities, concentration deficit, difficulties in learning through conventional teaching methods, disorganization, limited self-reliance and confidence, and low self-esteem.

Early intervention with appropriate teaching and assessment methods tailored to these children's needs can decrease their delays in learning and break their isolation. The interactivity between teacher and student is important for learning but is almost absent in the learning sessions due to the children's lack of understanding of the concepts being taught, and to the teachers' lack of media resources in remedial teaching. Multimedia technology can improve language learning for children with LD as we have shown in our previous work [4, 6–8]. In fact, multimedia can demonstrate new concepts using multiple appropriate modalities such visual, haptics, auditory, and gestures [9, 10]. It keeps the children engaged for a longer time and takes into consideration their different levels of difficulties to learn new concepts. Children can see a computer-animated tutor and can hear also the proper pronunciation of vocabulary [11]. Many research studies have shown that animations can improve the perception of new concepts especially in language learning [9, 10, 12, 13]. Learners who have used multimedia animations showed better progress and very good improvement in communication [5, 14–16].

A good number of educational systems exist in the literature. Cheng et al. [17] designed an online learning system for the Arabic language with ready-made content. Their system can select the learning material for every learner based on her knowledge that she provides at start. Erradi et al. proposed a simple game-like system called "ArabicTutor" [7] to teach Arabic. The content shows an Arabic word with its synonyms in different contexts with images. Wastam et al. [8] have developed a system to teach children stories through flashcards. The instructor selects a story and then asks the child to arrange its flashcards in a logical sequence. Rosmani and Abdul Wahab [18] proposed a simple prototype called "i-IQRA" to teach children the Holy Qura'n. It helps them to pronounce the verses in the right manner. Cheng et al. [19] proposed Crome for children and adults. Every lesson is followed by a set of exercises to assess the progress of the learners. Tabot and Hamada [20] have proposed a web-based educational system to teach physics to students. Ping et al. [21] have built an educational system for children with hearing impairment to teach them the Malay language. Wuang et al. [22] have built a multimedia courseware system that is based on learning theories. All these systems are based on static contents and address different learning objectives. However, none of the related work that we have seen so far could address all aspects related to Arabic language learning process (reading, writing, and observing and understanding) as the MOALEM platform is offering.

3 MOALEM Platform

We propose a new platform to improve the teaching of the Arabic language to Qatari children with LD and for nonnative Arabic learners at Qatar University. It is known that the learning process consists of reading, understanding, listening, and writing. However, many children cannot properly keep up with teachers in the classrooms and understand the Arabic vocabulary and grammar and pronounce words correctly. They need additional resources, personalized contents, and dedicated and skilled teachers. Very few elementary Qatari schools and educational centers have adequate staff and resources to help children with LD. For instance, the Shafallah center has different schools to teach children with different degrees of disability (i.e., mild, moderate, severe). However, the capacity of the center is limited and hundreds of children are on waiting lists. In addition, the number of children with LD is increasing every year in Qatar as the number of inhabitants is also growing considerably with the arrivals of more than 30,000 newcomers to work every month. Even though some schools have excellent teachers (e.g., Al Bayan schools), it is not, however, feasible for them to teach students in a one-on-one manner. Children with LD will have to seek other alternatives for learning. Parents will be obliged to hire private teachers to help their children keep up, which adds to the children’s expensive schooling. Figure 1 gives an overview of the MOALEM platform.

3.1 Dynamic Multimedia-Based Tutorial

This component of the platform consists of generating multimedia-based tutorials by mining Arabic text. It uses a core multidomain ontology that exploits existing Arabic natural language processing technologies including morphological analysis, logical rules, and discretization. The MOALEM platform will be able to understand Arabic story texts for children and generate personalized multimedia and adaptive tutorials. We develop an educational ontological model enhanced with an Arabic corpus that groups the terms, their synonyms, and multimedia elements. All the

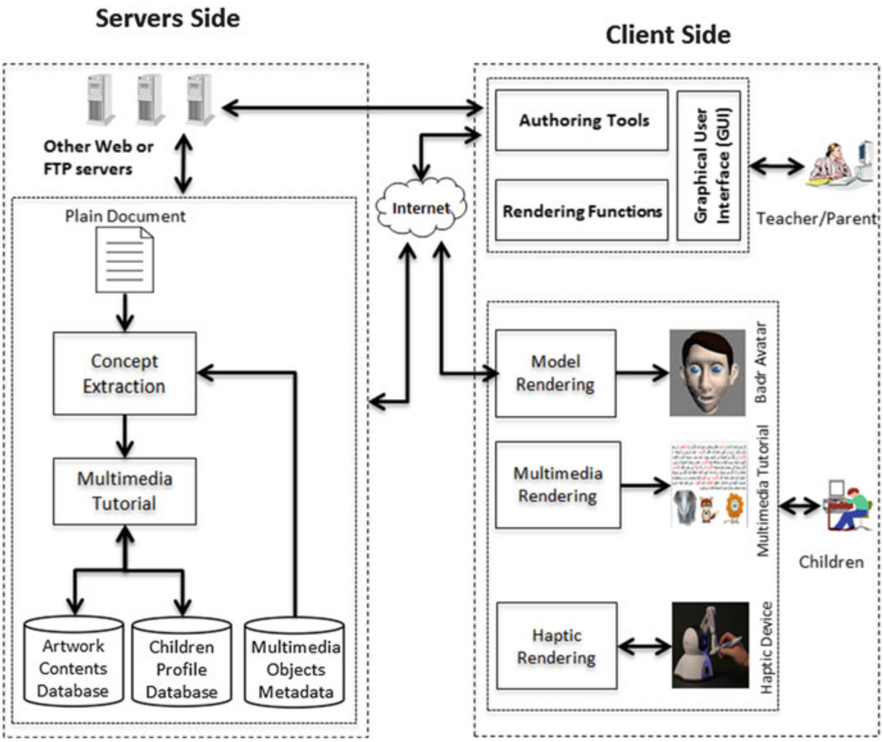


Fig. 1 The MOALEM platform architecture

concepts of the ontology are semantically linked. The instructors can then address semantic queries to the ontology and get multimedia-based responses. We use search engines (e.g., Google, and Yahoo) to get additional multimedia contents whenever needed to complement the explanation of the text. This component consists of Arabic text processing, educational ontology construction, tutorial generation, and assessment and evaluation. For morphological and orthographic disambiguation, we use MADAMIRA [23].

3.1.1 Arabic Text Processing

Arabic has a high degree of syntactic freedom that is attributed mostly to two phenomena: verb position alternations and case endings. The verb in Arabic often occurs at the beginning of a sentence, but it can also appear after the subject. Arabic nouns have case endings that allow some degree of freedom especially in poetic form. Given the above, the task of Arabic text processing must first fully orthographically and morphologically disambiguate the text as well as produce a

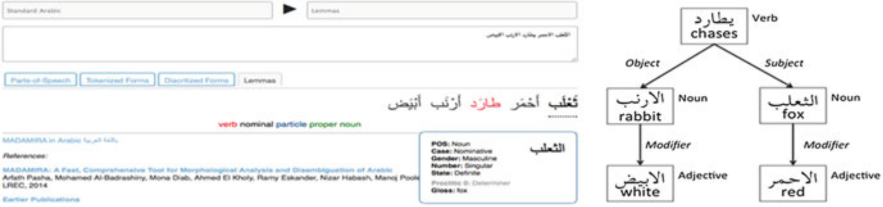


Fig. 2 Madamira Arabic morphology analyzer (left); CATiB Arabic parser output (right)

syntactic parse of it. Once done, an optimal set of linguistic features will be used in the MOALEM platform.

For morphological and orthographic disambiguation, we use the MADAMIRA system. MADAMIRA utilizes a rule-based morphological analyzer for words out of context and a set of classifiers trained on a corpus of Arabic annotated morphologically in context. The two resources are used to select for each word in context all of its morphological and orthographic features. These include the full diacritization of the word, identification of its stem and morphemes (clitics and affixes), part-of-speech, identification of the lemma (or citation form) of the word and its English gloss as well. The performance of MADAMIRA is quite competitive, scoring over 96% accuracy for lemma and stem as well as over 99% accuracy for word segmentation. Diacritization is also high (96%) when excluding the case markers (on which the performance drops to 86%). Finally, MADAMIRA's speed is close to 1000 words per second in a server–client mode. Figure 2 (left) shows the MADAMIRA online interface performance on the sentence الارنب يطارد الاحمر الثعلب الابيض " the red fox chases the white rabbit."

For syntactic analysis, we use the CATIB dependency parser [24]. The parser produces a simplified dependency representation, which is based on the Columbia Arabic Treebank (CATIB). Dependency representations abstract away from surface order, and thus allow the different verb positions in Arabic to be represented in the same way. For instance, Fig. 2 (right) illustrates the parse tree associated with the sentence discussed above regardless of whether the verb is sentence initial or sentence medial. The parser expects the input to be tokenized in the Arabic Treebank and Columbia Arabic Treebank tokenization scheme, and to also be part-of-speech tagged. This step must follow the initial morphological analysis and disambiguation step done by MADAMIRA. The parser's accuracy is almost 82% in terms of labeled attachment.

3.1.2 Educational Ontology Construction

An ontology is used to define the concepts of a domain and link them semantically. It is used for information modeling, sharing, and retrieval. Several domain-based ontologies have been recently developed. For instance, SNOMED ontology covers the medicine domain, and UNSPC ontology covers the products and services domain. We use an iterative approach that consists of designing first the global structure of the ontology, building a hierarchical taxonomy, design and fill the slots with instances, and finally validate the ontology. We create a knowledge base (KB) on the educational domain. This KB is enriched with the terms of new stories. The terms of interest are nouns (i.e., named entities, objects such as lion, dog, and cat), adjectives (color, size, etc.), and verbs (i.e., run, eat). We use an ontology of animal classes, as this is the most attractive domain for children. For instance, “Where does the camel live?”, “Which animals live with the camel?”, “Is the camel a carnivore or herbivore?” and so on.

3.1.3 Multimedia Tutorial Generation

For multimedia generation, we build a mapping component to link the dependency graphs with our ontologies according to each topic. The extracted relationships and entities are semantically mapped and validated according to the domain of discourse. Therefore, we need first to determine the similarity rating between the text-extracted key words, using the following formula that we adapted in our previous work [3]:

$$fSimilarity\,(zi, zj) = \begin{cases} w_{ls} fsim_{ls}\,(zi, zj) + w_{ss} fsim_{ss}\,(zi, zj) \rightarrow zi = zj \\ \\ fsim_{ls}\,(zi, zj) \rightarrow \quad zi \neq zj \end{cases}$$

Once we find an instance that matches the subject and the object, we can then search for a matching property for the predicate in the domain ontology model. SPARQL is used as a query language associated with our ontology. It is used to extract knowledge and infer new knowledge. When a SPARQL query is executed, a list of instances satisfying the request will be generated (e.g., “Lion” instance with all its details). Finally, we check if the concepts to which the instances for subject and predicate are asserted in domain and range of the property coincide. After generating the mapping, we combine them into search engine queries (i.e., Google search query) in order to retrieve the corresponding multimedia-based which then it will be semiautomatically ranked and presented to the teacher. The proposed approach allows us to get detailed information from the Arabic story text including the following: character names, actions, and events. We send logical queries in simple Arabic words to the ontology to retrieve the corresponding multimedia elements. Figure 3 shows the result of processing an Arabic sentence and its conversion to multimedia elements. All the previous Arabic text disambiguation techniques can be used to improve the accuracy of text mining.

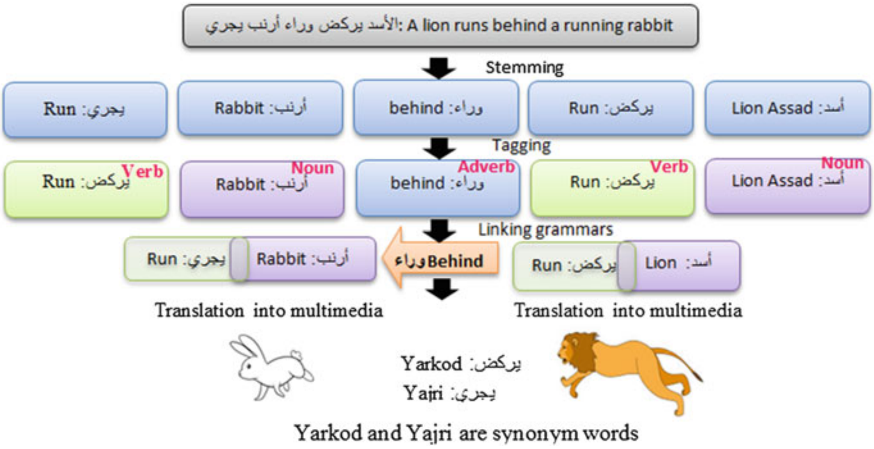


Fig. 3 An Arabic sentence processing and mapping to multimedia

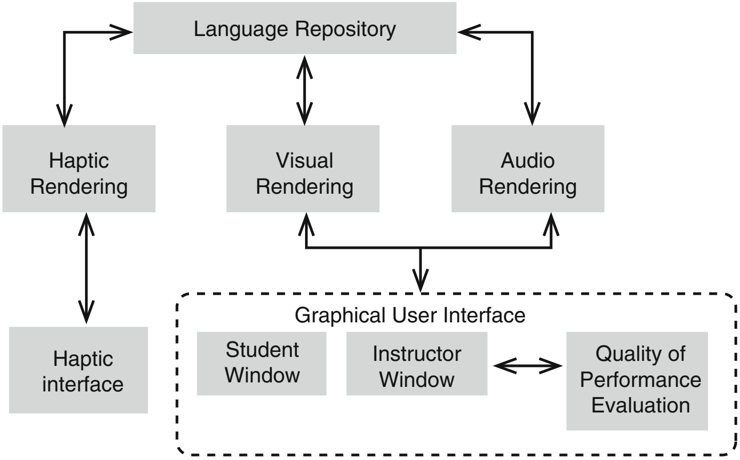


Fig. 4 ALKATIB system architecture

3.2 *ALKATIB: The Haptic Handwriting Component*

A simplified architecture of the ALKATIB system is shown in Fig. 4. It comprises six modules: (1) a language repository (storing Arabic alphabet data), (2) haptic rendering, (3) audio and visual rendering, (4) haptic interface, (5) a Quality of Performance Evaluation module, and (6) a Graphical User Interface (GUI). Note that the haptic system setup costs less than US$300 ($250 for the Novint Falcon device and less than $50 for the custom grip and software).

The proposed design for the haptic device is shown in Fig. 5. The Graphical User Interface for the ALKATIB system is made up of two windows: instructor window

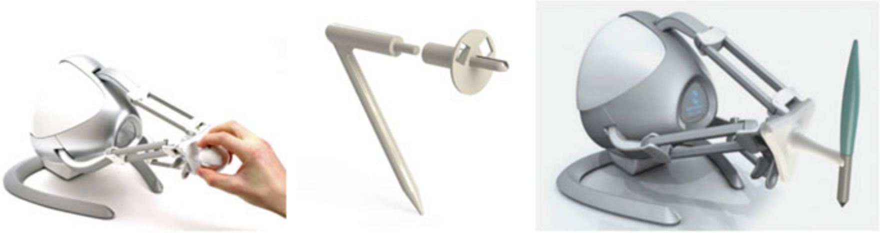


Fig. 5 The proposed design for the haptic handwriting device

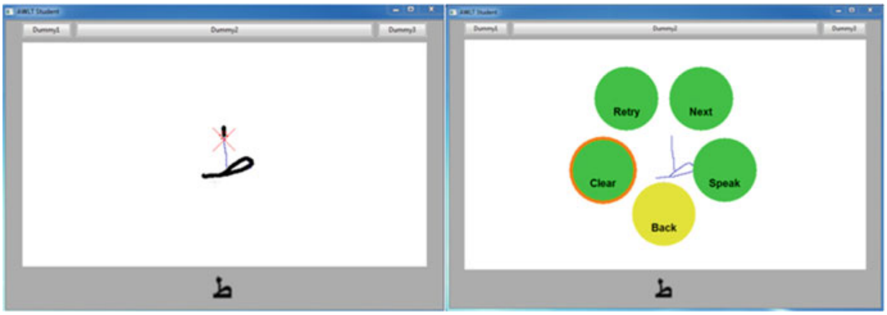


Fig. 6 The proposed system interface

and the student window. The student window (Fig. 6) enables the learner to load and play back handwriting tasks. The instructor window enables the instructor to author handwriting tasks by recording haptic, audio, and visual contents and display them to the student. The instructor window also provides a control panel to customize the haptic and the visual rendering to suit specific needs of the students.

3.2.1 Performance Evaluation for ALKATIB

The objective of the experiment is to compare partial haptic guidance and full haptic guidance to improve learning outcomes. The experimental setup included a laptop, the haptic interface (Novint Falcon haptic device with the custom grip), and the software application running on the laptop. The laptop has an Intel Core i7-2640 M CPU running at 2.80 GHz, 8 GB of RAM, an Intel HD Graphics 3000, and runs Windows 7 professional operating system (64-bit). A snapshot of the experimental setup is shown in Fig. 7. A total number of 22 adult users participated in the experiment who were divided into two groups, each one consisting of five females and six males. The age range was 18–45 years.

Group 1 began its training with the full haptic guidance mode in the first three sessions and then moved on using the partial haptic guidance in the last three sessions. Group 2, on the other hand, started with the partial (first three sessions)

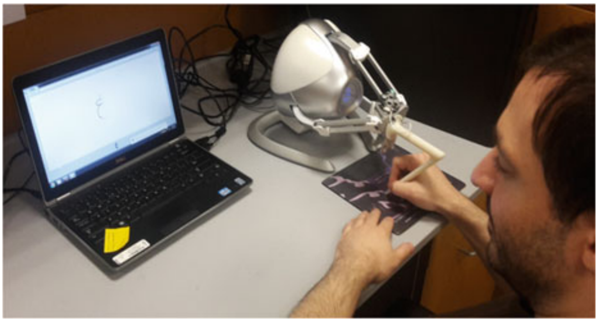


Fig. 7 The experimental setup for ALKATIB

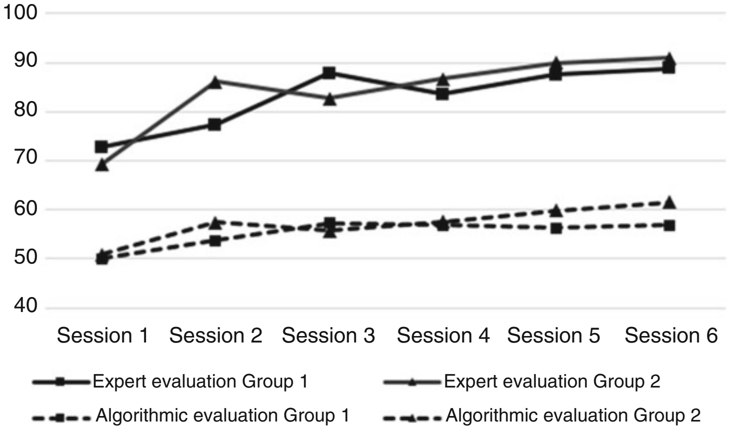


Fig. 8 Expert (top) and algorithmic (bottom) evaluations per group per session

and continued with the full haptic guidance mode (three last sessions). We compare the average scores the participants earned by the end of the first session and the end of the last session for both expert and algorithmic evaluations. These results are depicted in Fig. 8. The improvement in the average score for Group 2 (21.5%) is significantly higher than the improvement in the average score for Group 1 (16.1%). The same conclusion can be derived by examining Fig. 9 (left) (algorithmic evaluation) and Fig. 9 (right) (expert evaluation).

Comparing partial haptic guidance and full haptic guidance, it seems that when learning the gross aspects of handwriting trajectory, partial guidance is more efficient, while learning fine details of the handwriting is conveyed better with full guidance. This suggests that learning generic handwriting skills may utilize partial haptic guidance, whereas personalized handwriting skills can be learned better through full guidance.

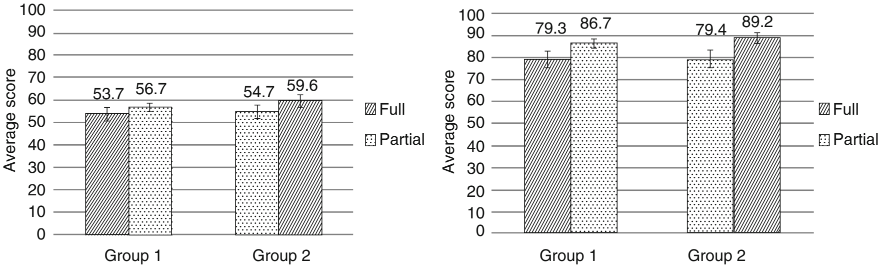


Fig. 9 Algorithmic evaluation (left); expert evaluation (right)

3.3 ALNATEQ Badr: The Speech Synthesis Component

Research has shown that the perception of a language vocabulary is influenced by the speaker's face, gestures, and sounds [25]. These elements are particularly valuable for children with language learning difficulties. Therefore, we propose to create a realistic animated talking face "Badr" and use it as a tutor in language learning with the children. Creating a realistic Arabic animated talking face will involve the following tasks. The computer-animated face is critically dependent on a speech synthesizer that provides phoneme and duration information as well as auditory speech. We create an Arabic auditory speech synthesizer and validate it with the teachers. Speech synthesizers put all of a language's basic segments in memory and then combine these appropriately to generate new speech [26]. Thus, any new text can be used as input. The written text is translated into a phonemic representation and the segments in memory are optimally chosen and concatenated to provide the auditory speech. The computer-animated face uses the phoneme representation and the durations of the phonemes to create the appropriate facial and tongue animation, which is appropriately aligned with the synthesized auditory speech. MOALEM accesses the auditory speech database in order to carry out auditory speech synthesis.

Badr currently is implemented on a PC for development and as an application on iPhone devices. We will design an Arabic looking talking face, animate it with respect to Arab cultural behaviors and norms, and integrate it into the MOALEM platform window to make the talking face readily available to the teacher. Previous research has also developed computer-animated talkers in a variety of languages. Most relevant to the present proposal is the development of an Arabic talking face, which has been shown to be extremely accurate and to be effective in language learning. We will adapt our computer-animated tutor Baldi to pronounce Arabic words in standard Arabic. The modifications will be made on Baldi's control parameters of the polygon model. One set of parameters controls the movement of vertices and their immediate neighbors. Geometric changes include rotation such as jaw rotation or translation in location of the vertices such as mouth widening. The scale and subareas of the face can also be changed such as in the cheek. The effect of the face can also be changed (e.g., showing happiness or sadness) by control parameters.

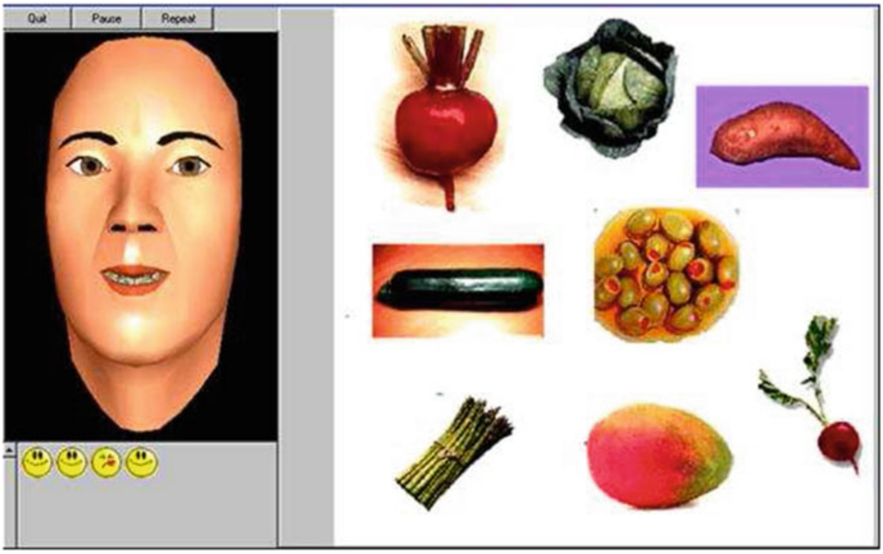


Fig. 10 The talking face Badr

Animation of the talking face Badr synthesizes a sequence of phonemes. Using text-to-speech synthesis, a written utterance is mapped into a phoneme sequence and their corresponding control parameters. Rounding and jaw rotation are examples of control parameters. The control of Badr also implements coarticulation, which means that the nature of a phoneme is influenced by surrounding phonemes. Direct instruction is an effective method for teaching new vocabulary [27]. Implementing different forms and representations of class material has tremendous impact on learning. Text along with visual images can be paired with appropriate definitions as well as the speech of the words to be learned. This multisensory approach with an English computer-animated agent has been effective in several studies with deaf and hard-of-hearing children [1]. Many new words and grammar were mastered in peer-reviewed experiments [28, 29]. Figure 10 shows a lesson to teach fruits and vegetables. Each lesson includes the talking face, the vocabulary words, and "stickers." The goal is for the child to learn the most used vocabulary of vegetables and fruits. Badr might say "click on the olives." If the child selects the right item, he is rewarded with praise and a happy face.

4 Evaluation and Assessment

Two groups of 10 children each are identified to use the platform: the first one will use MOALEM platform with its components while the other group will continue learning through the current conventional teaching methods. We will conduct the assessments periodically with the two groups measuring their advancement skills in reading, understanding, and writing of Arabic vocabulary and sentences. A set of specific words will be determined and used in both groups. Our assessments include several measurements and metrics that we have developed and used in our previous research [2, 4] that include time of reading a paragraph, explanation of the words, time of writing, and time to retrieve the meaning of words. Other metrics proposed in our work [30] are applied. All potential users of the system, including teachers, children, parents, and caregivers, will be requested to provide feedback on the benefits of the platform, its interfaces, effectiveness, usage, generation of adaptive and personalized learning tutorials, problems, and suggestions for improvements. In order to assess the reading/comprehension outcomes, we will utilize the Neuropsychological Assessment Battery for children. Nonverbal IQ skills will be measured with WISC nonverbal IQ subsets including picture completion, picture arrangement, block design, object assembly, digit-coding symbol, and mazes subtests. Furthermore, two Rapid Automatized Naming (RAN) tasks will be administered to the children (a picture RAN and a digit RAN). Finally, Arabic texts will be created for the purpose of the experiment (one fully vowelized and one non-vowelized). As for handwriting, the Evaluation Tool of Children's Handwriting (ETCH) will be used to assess the quality of typical children handwriting. For the clinical sample, assessment will be done using the Wechsler Preschool and Primary Scale of Intelligence[3] (WPPSI-III) for 5-year-old patients, and the Wechsler Intelligence Scale for Children (WISC-IV) for patients aged 6 and above. The cutoff range for selection will be 50–79, that is, mild intellectual difficulty to borderline difficulty range. The assessment will be divided into two parts. Part 1 will assess individual learning elements of the MOALEM platform (reading, writing, or speaking). We implement the multisensory computer-controlled environment to teach vocabulary and grammar directly. To test whether our computer-animated tutor, Badr, is responsible for the learning of new vocabulary, we carry out experiments using a multiple baseline design for each child. In this design, all words are tested, whereas a subset of these words is being trained as well as tested. Using this procedure guarantees that successful learning of the words only during training means that the training regimen was responsible for the learning. The lessons contain vocabulary words that are unique for each child. Each lesson consists of 24 words, broken down into three subsets of eight words each. The Arabic computer-animated talker says the word when images of the vocabulary items are selected, as shown in Fig. 4. The child responds to Badr's instructions such as "click on the olive," or "show the beets." Reading exercises allow the child to recognize and type the word. Badr will also have the child pronounce the word when Badr names a highlighted image or simply highlights the image without pronouncing it.

[3] http://link.springer.com/referenceworkentry/10.1007%2F978-0-387-79948-3_1606

5 Conclusion

We proposed a new platform to teach children with learning difficulties the Arabic vocabulary using multimedia technology. The platform supports reading, writing, and listening of Arabic words. It has a customized haptic device for writing and a talking face to pronounce Arabic vocabulary with gestures. The platform can be used in schools' settings as well as at home where parents can assist their children to review the materials they study in schools. We have developed an educational ontology that allows the instructor to get semantically related information about words.

Acknowledgment This work was made possible by NPRP grant #10-0205-170346 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Massaro, D.W., and Light, J.: Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals," J. of Speech, Language, and Research, 47(2), pp. 304–320, (2004).
2. Saleh, M., and Alja'am J.M.: A Fully Accessible Arabic Learning Platform for Assisting Children with Intellectual Challenges, in 14th International Conference on Computers Helping People with Special Needs, Lectures Notes in Computer Sciences, pp. 1–8, Paris, France (2014).
3. Dandashi, A., Saad, S., Alja'am, J.M., Saleh, M.: The Multimedia-based Learning System Improved Cognitive Skills and Motivation of Disabled Children with a Very High Rate, in Journal of Educational Technology & Society, pp. 1–8, (2014).
4. Karkar, A.G., Saleh, M., Saad, S., Al Ja'am, J.M.: An Arabic Ontology-based Learning System for Children with Intellectual Challenges, IEEE Global Engineering Education Conference, pp. 670–675, (2014).
5. Moreno, R. and Mayer, R.E.: Engaging Students in Active Learning: The Case for Personalized Multimedia Messages, in Journal of Educational Psychology and Technology 92: pp. 724–733, (2000).
6. Saleh, M.S., Alja'am, J.M., Karime, A., El Saddik, A.: MeMaPad: An Edutainment System for Assisting Children with Moderate Intellectual and Learning Disability. Case Study: Children at the Shafallah Center in Doha, Qatar, In Proc. of the International Conference on Technology for Helping People with Special Needs, Riyadh, Saudi Arabia, pp. 69–74, (2013).
7. Erradi, A., et al.: ArabicTutor: A multimedia m-Learning Platform for Learning Arabic Spelling and Vocabulary, in the IEEE International Con. on Multimedia Computing and Systems, pp. 833–838, (2012).
8. Wastam, Jumail, et al. : A Guided Digital Storytelling Prototype System using Illustrated Flashcards, pp. 1–6, (2010).
9. Liu, M., Z. Moore, L. Graham and Lee, S.: A Look at the Research on Computer-Based Technology Use in Second Language Learning: A Review of the Literature from 1990–2000, in Journal of Research on Technology in Education 34(3), pp. 250–273, (2002).
10. Liu, M.: Hypermedia Assisted Instruction and Second Language Learning: A Semantic-Network-based Approach, in Computers in School 10(3–4), pp. 293–312, (1994).

11. Wik, P. and Hjalmarsson, A.: Embodied Conversational Agents in Computer Assisted Language Learning. Speech Communication, 51(10), pp. 1024–1037, (2009).
12. Park, N. and Son, J.: (2009): Implementing Computer-assisted Language Learning in the EFL Classroom: Teachers Perceptions and Perspectives, in Journal of Pedagogies and Learning 5(2): pp. 80–101, (2009).
13. Lee Swanson, H., Karen R. Harris, Graham, S.: Handbook of Learning Disabilities, Second Edition, ISBN-13: 978–1462518685, (2014).
14. Eid M., M. Orozco, and El Saddik, A.: "A Guided Tour in Haptic Audio Visual Environments and Applications", Int. Journal Adv. Media Communication, vol. 1, issue 3, pp. 265–297, (2007).
15. Senatore, R. and Marcelli, A.: A Neural Scheme for Procedural Motor Learning of Handwriting, In International Conf. on Frontiers in Handwriting Recognition, (2012).
16. Beck, I. L., McKeown, M. G., and Kucan, L.: Bringing Words to Life: Robust Vocabulary Instruction, New York: Guilford Press, (2002).
17. Cheng, I, Anup Basu, and Goebel, R.: Interactive Multimedia for Adaptive Online Education, In IEEE Multimedia, vol. 16, no. 1, pp. 16–25, (2006).
18. Rosmani, Arifah Fasha, and Abdul Wahab, N.: "i-IQRA": Designing and Constructing a Persuasive Multimedia Application to Learn Arabic Characters, in the 2011 IEEE Colloquium on Humanities, Science and Engineering, pp. 98–101, (2011).
19. Cheng, I, Basu, A., and Goebel, R.: "Interactive Multimedia for Adaptive Online Eeducation," In the IEEE Multimedia, pp. 16–25, (2009).
20. Tabot, A., and Hamada, M.: A Multimedia Learning System for Selected Topics of Physics, in the IEEE International Conf. on Information Technology Based Higher Education and Training, pp. 1–8, (2012).
21. Ping, T. P., Sharbini, H., Chan, C. P., and Julaihi, A.A.: Integration of Cultural Dimensions into Software Localisation Testing of Assistive Technology for Deaf Children, In 5th Malaysian Conference on Software Engineering, pp. 136–140, (2011).
22. Wuang, P., Chiang, S., Su, Y., and Wang, C.: Effectiveness of Virtual Reality using Wii Gaming Technology in Children with Down Syndrome, R. in Developmental Disabilities, 32(1), pp. 312–321, (2011).
23. Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth (2014): "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," In Proceedings of the Language Resources and Evaluation Conference, Reykjavik, Iceland.
24. Marton, Yuval, Nizar Habash and Owen Rambow (2013): "Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features," Computational Linguistics, Vol. 39, Issue: 1.
25. Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J., and Clark, R.: Animated Speech: Research Progress and Applications. In G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds.) Audiovisual Speech Processing, pp. 309–345. Cambridge University Press, (2012).
26. Massaro, D.W., Bosseler, A., and Light, J.: Development and Evaluation of a Computer-Animated Tutor for Language and Vocabulary Learning, In Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2003.
27. Massaro, D.W., Ouni, S., Cohen, M.M., and Clark, R.: A Multilingual Embodied Condversational Agent, In R.H. Sprague (Ed.), Proceedings of 38th Annual Hawaii Int. Conference on System Sciences, Los Alimitos, CA: IEEE Computer Society Press, (2005).
28. Massaro, D. W., and Light: Read my Tongue Movements: Bimodal Learning to Perceive and Produce Non-Native speech /r/ and /l/. Eurospeech 2003-Switzerland (In-terspeech). 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, (2003).

29. Chen, T. H. and Massaro, D.W.: Evaluation of Synthetic and Natural Mandarin Visual Speech: Initial Consonants, Single Vowels, and Syllables, Speech Communication 53, pp. 955–972, (2011).
30. Duff, S., and Goyen, T.A.: Reliability and Validity of the Evaluation Tool of Children's Handwriting–Cursive (ETCH–C) using the General Scoring Criteria. American Journal of Occupational Therapy, 64, pp. 37–46, (2010).

Person-Dependent and Person-Independent Arabic Speech Recognition System

Noor Al-Maadeed and Somaya Al-Maadeed

1 Introduction

Conversational interfaces have greatly improved over the last 20 years, leading to a new era of human–computer interaction. Voice recognition has recently begun to play an important role in information retrieval methods. However, while voice recognition technology for English speakers has been under investigation for more than 30 years, only a few studies have been performed on Arabic sounds.

Sounds can typically be recognized through the unique characteristics of each word. However, the pronunciation of a given word varies in a complex fashion among individuals, similar to music. Every person's voice is unique, and the same words can have different meanings depending on tone or perspective. Several approaches have been taken to solve this problem.

There are two popular frameworks for speech recognition: hidden Markov models (HMMs) [1], and artificial neural networks (ANNs) [2]. HMMs are simpler, faster, and generally require less training data than ANNs for most recognition tasks. In addition, HMMs can characterize the speech signal in a mathematically tractable fashion [3]. A hybrid HMM–ANN has recently been proposed to improve the performance by combining the two modeling strategies [3, 4].

Template matching is often used in audio classification problems. However, straightforward template matching often fails [5]. In [6, 7], continuous Gaussian mixture density models (CDHMMs) are shown to have error rates at least comparable to the rates of the best template recognizers and significantly lower than the

N. Al-Maadeed · S. Al-Maadeed (✉)
Department of Computer Science and Engineering, Qatar University, Doha, Qatar
e-mail: n.alali@qu.edu.qa; s_alali@qu.edu.qa

J. M. Alja'am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_16

error rates of discrete symbol HMMs. This work focuses on the CDHMM approach because of its proven reliability and clear formulation.

In this chapter, the Arabic spoken word recognition problem is modeled using an HMM. The HMM states are identified with the sounds of the letters of the alphabet. Once the model is established, the Viterbi algorithm is used to recognize the sequence of letters composing the word.

2 Unique Characteristics of Arabic Words

The attributes, functionalities, and limitations of Arabic (or other non-Latin) speech recognition have been investigated previously. However, little attention has been devoted to the specific characteristics of Arabic word pronunciation. “Arabic is a Semitic language with approximately 221 million speakers in the Arab world and some African and Asian countries such as Chad, Cyprus, Iran, Israel, Kenya, Mali, Niger, Tajikistan, Tanzania, etc. ...” [8]. In addition, there are over 30 different varieties of colloquial Arabic.

English words are dissimilar to Arabic words in several ways, the most obvious and significant being the alphabet and symbols. There is no IPA (International Phonetic Alphabet) for the pronunciation of Arabic letters such as ع, ص، ض، ظ، ط and خ. Alghamdi Mansour (Mansour) has introduced a solution to this problem. Another distinct characteristic of Arabic words is their rhythm. The vowel and consonant lengths can affect the meanings of Arabic words, which is not the case in English. Even the most accurate voice recognition application does not reach 100% accuracy in Arabic. In this chapter, we study the ten Arabic numbersسبعة ثمانية, ستة, خمسة, ثلاثة اربعة, اثنين, واحد تسعة and وعشرة which include the unique Arabic letters ح, ع, خ ث. For multiple users, we studied the most popular Arabic words ان, من, في ما, مع, عن, الذي, التي, الى, على, most of which contain the letter ع. We then examined the most frequently used English words (“the,” “of,” “and,” “in,” “to,” “was,” “it,” “is,” “for,” and “that”), omitting one-letter words such as “I.”

3 System Overview

Our speech recognition system operates in three stages: feature extraction, generating the HMM model, and testing the model. The underlying model was a word-based HMM, as illustrated in Fig. 1. The following sections describe the training and testing of the model, including a discussion of the sound features used in the system. The HMM classifier, which classifies the features captured from the word image, is discussed in Sect. 4. The experimental results are presented in Sect. 5.

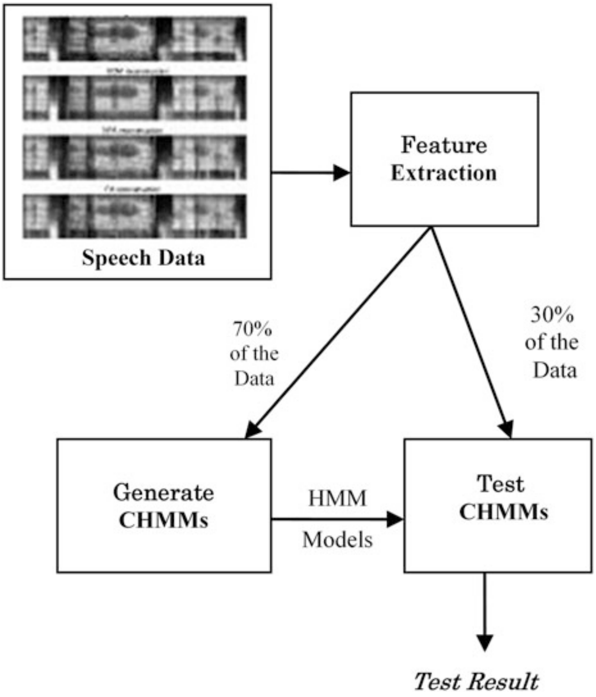


Fig. 1 The Arabic speech recognition procedure using the CDHMM

4 Data Content Preparation

In this stage, all of the data required by the system is collected, arranged in a suitable format for later use and stored. Three ten-word vocabularies are considered. The first vocabulary consists of the ten Arabic digits (TAD) between one and ten, as shown in Table 1. A word recognition test is performed by collecting the word vocabulary from ten different speakers. The Arabic numbers are used because many existing databases include a spoken digit. We recorded 100 wave sound files for ten different people, each of whom were asked to repeat each Arabic digit ten times.

The second vocabulary consists of the ten most popular Arabic words (MPAW), a list selected based on a prior study in which the number of occurrences of 30,000 words was counted on various Arabic websites. The top ten most common words (على, ان, من, في م, مع, عن, الذي, التي, الى ,) were selected. The voice of a single speaker was recorded ten times for each of these ten words.

The third ten-word vocabulary consists of the most used English words (MUEW) based on the whole corpus (spoken and written English): the, of, than, in, to, was, it, is, to, and for. The same speaker voice was recorded ten times for each of the ten most popular words in Arabic and English.

Table 1 The ten Arabic digits (TAD) between one and ten, with their meanings in English and transliterations [9]

	Arabic word	Meaning in English	Pronunciation
1	واحد	One	*Waahed*
2	اثنين	Two	*Athneien or athnan*
3	ثلاثة	Three	*Thalatha*
4	اربعة	Four	*Arbaa*
5	خمسة	Five	*Kamsaa*
6	ستة	Six	*Sitaah*
7	سبعة	Seven	*Saba'a*
8	ثمانية	Eight	*Thmaniea*
9	تسعة	Nine	*Tisa'a*
10	عشره	Ten	*A'ashra*

5 Feature Extraction

The feature extraction procedure consists of converting the voice sample into a series of appropriate vectors that describe the "features" of the signal. Feature extraction is performed regularly via short-term spectral analysis, which divides the speech signal into short frames of fixed length that can be assumed to be independent over a short time interval. Each frame usually overlaps its neighbor. To avoid "spectral artifacts" caused by discontinuities at the frame endpoints, the data in each frame is convolved with a smooth window function. The Hamming window is the most widely used smoothing function in speech recognition.

In most speech recognition systems, the acoustic features of choice are the Mel Frequency Cepstral Coefficients (MFCCs) [10]. The MFCCs are robust, contain a wealth of information regarding the vocal tract configuration regardless of the source of the excitation, and can be used to represent all classes of speech sounds. Other features, such as the Perceptual Linear Predictive (PLP) [11] coefficients, may also be used. In [12], an interesting set of acoustic parameters was presented, and their association with specific phonetic features was thoroughly investigated. However, the implementations presented in [12] are generally based on MFCCs.

6 The HMM Classifier

A Markov model can be described as a system with a set of states and a set of transitions between states. Each transition has an associated probability, and the system proceeds from state to state based on the current state and the probability of the transition to a new state. The observability of the states determines whether a Markov model is hidden. In standard Markov models, the states are directly observable. HMMs have states that are not directly observable; instead, there is a set of possible observations for each state and a set of allowed state transitions, and the observations of any one state depend on the probabilities of the possible observations.

When HMMs are used in pattern recognition, a set of observations is provided as an input signal. The HMM classifier then attempts to decide which of a set of possible HMMs is most likely to generate the given set of observations. The classification system contains a number of HMMs, each corresponding to a category, and the class corresponding to the HMM that best reproduces the incoming signal is the category assigned to that signal. HMMs work well with sounds that vary in duration because the durational change can occur in a single state or across many states. An HMM can begin by following the signal state for state, jump back and forth between a few middle states for the duration of the sustained portion of the sound, and then follow the decay portion down state for state. This procedure models the sound more accurately than a template that must be stretched to fit the length of the input signal.

An HMM is a stochastic process with an underlying Markov process that can only be observed through another set of stochastic processes produced by the Markov process (the observations are probabilistic functions of the states). Let us assume that a sequence of observations, $O = (o1, \ldots, oT)$, is produced by the state sequence $Q = (q1, \ldots, qT)$, where each observation ot is from the set of M observation symbols $V = \{vk;\ 1 \leq k \leq M\}$, and each state qt is from the set of N states $S = \{si;\ 1 \leq i \leq N\}$. An HMM can be characterized by $\pi = \{\pi i\}$, where $\pi i = P(q1 = si)$ is the initial state probability; $\pi = \{a_{ij}\}$, $aij = P(qt + 1 = sj \mid qt = si)$ is the state transition probability; and $\pi = \{bj\,(k)\}$, where $bj\,(k) = P(ot = vt \mid qt = j)$ is the symbol probability. The following probability constraints (1) must be satisfied:

$$\sum_{i=1}^{N} \pi = 1; \quad \sum_{j=1}^{N} a_{ij} = 1 \forall i; \quad \sum_{k=1}^{M} b_k(b) = 1 \forall j. \tag{1}$$

The HMM is specified using the compact notation $\lambda = \{\Pi, A, B\}$. We have only 55 classes (rather than 120 classes corresponding to the set of Arabic characters) because the recognition system is implemented on a limited set of words (the Arabic numbers). The 55 letters or sub-letters of the alphabet are defined to be the states of our HMM, where the initial πi and the transition probabilities a_{ij} are computed as follows:

$$a_{ij} = \frac{\text{number of trans.from } l\,(q_i)\ \ to\ l\left(q_j\right)}{\text{number of transitions from } l\,(q_i)} \tag{2}$$

where the function l maps a state to its representative member of the alphabet. Our system has two phases, training and testing, as illustrated in Fig. 1.

A. Training problem

Given the training sequence $O = o_1, \ldots, o_T$, the model parameters, $\lambda = \{\pi, A, B\}$, were adjusted to maximize $P(O/\lambda)$. The Baum–Welch Algorithm was used to determine the optimization criterion for finding the Maximum Likelihood (ML). In general, the Baum–Welch algorithm provides satisfactory performance in the training of HMMs [13].

B. **Recognition phase**

A Modified Viterbi Algorithm (MVA) was used to solve the recognition problem in the recognition phase [14].

7 Experimental Results

This section analyzes the output of our system to determine how well it recognizes Arabic speech in multi-user and single-user environments. The performance in an Arabic environment is also compared to the performance in an English single-user environment. An evaluation of the continuous hidden Markov model implementation phase was performed to determine its score on Arabic words compared to English words. This section describes the results of these tests and the problems encountered. As discussed in Sect. 3, we study the recognition of the ten Arabic digits (TAD) for HMMs with different combinations of parameters in a multi-user environment and then for different users. The approach taken in this project is as follows: The system performance was evaluated based on two criteria: the recognition rate and how well each word is recognized compared to other possible incorrect solutions.

The same procedure was performed on the most popular Arabic words (MPAW) and the most used English words (MUEW). The recognition rate was calculated for each vocabulary, and the discrepancies among different possible solutions were analyzed. The results for the Arabic word list were then compared with the results for the English words.

The results can be divided into two main parts: the results for a multi-user environment and the results for individual users. The list of ten Arabic digits (TAD) is used for the first (multi-user) part of the analysis. Two aspects of the model were analyzed: first, the effect of the choice of model parameters (the number of states, number of mixtures, and number of iterations) and second, the word characteristics of the different models. The words were tested on different individuals and drawn if there was a character for the Arabic word.

The second part of the analysis involves individual user voice recognition for Arabic and English words. The system is tested on MPAW (the most popular Arabic words) and MUEW (the most used English words). Two quantities were calculated from the output conflict matrix described in Chap. 4: the recognition rate and the variation among the optimal solutions.

To determine the role of the individual character or other word in the evolution process and the system implementation, a wider-scale application of the system will be discussed. Suggestions for improving the evaluation and program are discussed at the end of this chapter along with the limitations of the project.

The recognition rate for the three databases is calculated using a combination of three variables: the number of states, the number of mixtures, and the number of iterations. This section discusses the TAD (ten Arabic digit database) performance in a multi-user environment. The details of the measurements and performance evaluations for multi-user TAD and for the individual-user MPAW and MUEW environments are discussed in the following subsections.

7.1 TAD Results for Multiple Users

The proposed system first loads files for each word from word "one" to word "ten." Ten samples are used for each speaker (for a total of 100 files) in the feature extraction to generate the HMMs, training model and test model. The overall recognition rate is shown in Table 2.

The CDHMM training model employed 70% of the voices, and the testing model used 30% of the voices. By analyzing the outcome of the test model, it was found that different results are obtained for different values of the variables (the number of states or number of iterations). The optimal number of states is seven or three, and the number of states yielding the worst performance is four. The number of iterations is optimal when the number of mixtures is equal to two or four and decreases as the more mixtures are added. This trend occurs because the training data for each HMM model consists of seven voices. Any number higher than seven is therefore inadmissible. The number of iterations decreases and then stabilizes to a constant value of one. All of the users are female, but two regional accents were detected. The speakers also vary in their degree of consistency. Some individuals are consistent in their speech; each time these individuals are asked to pronounce a word, they produce similar acoustic (acoustic) signals with only minor variations. Consistent speakers are referred to as "sheep," while inconsistent speakers are referred to as "goats" [14]. Speaker number 10 in the TAD sample was clearly a "goat" as the measurement for this speaker was always the closest to zero (had the smallest absolute value). The variance percentage in the solution for multiple users is −1.5.

The effects of individual variation were also studied by analyzing the HMM parameters for the multi-user Arabic dataset. Various sets of volunteers were examined and trained to determine how the differences among individual speakers might affect the results. In this analysis, the datasets used for the training and the speaker used in the model evaluation were varied. TAD1 was used as the training model for volunteers/speakers 1–7 and tested on the remaining volunteers

Table 2 Best recognition rates for a single user and for multiple users

Database	Recognition rate (%)	Variance measurement
TAD	84	−1.5
MPAW	83	−0.34
MUEW	76.6	−0.28

Table 3 Recognition rate of different users

Training data (model No)	Test no.	User providing test data	Recognition rate (%)	Required solutions variance
TAD1	1	User no. 8	50	−1.41
TAD1	2	User no. 9	70	−1.15
TAD1	3	User no. 10	70	−0.34
TAD2	4	User no. 1	80	−2.14
TAD2	5	User no. 2	90	−1.6
TAD2	6	User no. 3	90	−1.9
TAD3	7	User no. 3	70	−1.87
TAD3	8	User no. 6	90	−1.87
TAD3	9	User no. 9	80	−1.19
Average	10			−1.5

(8–10), while TAD2 was the training model for volunteers 4–10 and was tested on volunteers 1–3. TAD3 was also tested for every third volunteer in the row (volunteers 3, 6, 9) and trained on the remaining voices.

Table 3 demonstrates that different speakers produce different recognition rates. However, the recognition rates may also be affected by factors that have not been considered. For example, speakers 9 and 10 produce similar recognition rates (70%), but the other possible solution may approximate the required solution. In other words, some speakers may speak more clearly than others yet produce the same recognition rate. A measure of the deviation of the alternative solutions from the required one is therefore needed. A larger absolute value of the measure indicates a better result. Therefore, speaker 9 with TAD1 has a much clearer voice than speaker 10 with the same training data (TAD1), as shown by the measurement of 1.15 for speaker 9 compared to 0.34 for speaker 10.

Table 3 demonstrates that different speakers have different recognition rates. The highest recognition rates (90%) are obtained for speakers 2 and 3 with TAD2 and speaker 6 with TAD3. The lowest recognition rate (50%) was obtained for speaker 8. It has also been observed that a speaker may produce different recognition rates depending on the training data used. For example, the recognition rates for the third speaker vary from 90% (with the TAD2 model) to 70% (with TAD3); this means that the recognition rate depends not only on the tested user but also on the training voices. In addition, note that the pronunciation of the words is marginally similar (−1.87 and −1.9) and slightly above the average value, possibly because the speaker still speaks with the same clarity.

The same observation applies to speaker 9, who delivers a recognition rate of 70% and 80% using TAD1 and TAD3, respectively. In addition, the variance in the result is similar for both cases (−1.15 and −1.19), demonstrating that the tested voice outcome was not influenced by the training voice.

Table 3 also shows that speaker 8 produces a recognition rate of only 50%. Nevertheless, speaker 10 (with a recognition rate of 70%) may have inferior voice recognition compared to speaker 8 as the measurement in the case of a distinguishable solution is −0.34 compared to −1.4 for speaker 8 (see Table 4).

Table 4 Recognition rate for entire words (Number of states = 3, Number of mixtures = 3)

Data set	Average recognition rate (%)	Optimal solutions variance
TAD1	63	−1.11916
TAD2	76	−2.3454
TAD3	80	−2.45801

Table 5 Characteristics of each word in the TAD database

Word	Recognition rate (%)	Required solution variance × −20	Required solution variance
واحد	77.78	54.6	−2.73
اثنين	77.78	31	−1.55
ثلاثة	55.56	31.8	−1.59
أربعة	88.89	41.8	−2.09
خمسة	55.56	42.8	−2.14
ستة	55.56	50.2	−2.51
سبعة	88.89	44.6	−2.23
ثمانية	55.56	20.8	−1.04
تسعة	77.78	57.6	−2.88
عشرة	100	70.6	−3.53
All	73.33	30.2	−1.51

The third group (TAD3) delivers the highest recognition rate, while the recognition using the first group (TAD1) is lowest. This result occurs because speaker 9, with a poor recognition rate, was included in TAD1. The other groups achieve a high word recognition accuracy. The probability also increases as the voice recognition improves for the other groups. The words three and eight have low overall recognition rates, again because TAD1 includes an unclear speaker, especially for those two words. Both words include the letter "thaa ث."

Table 5 summarizes the recognition results for every word in the TAD database. The word ten عشرة is always recognized, while the words three ثلاثة, five خمسة, six ستة, and eight ثمانية have a much lower recognition rate (55%). The word eight ثمانية is the least distinguishable (the solution variance is −1.08), followed by the words three, five, and six. The words four "اربعة "and seven "سبعة "have high recognition rates. All of the words with recognition rates of 88% or higher include the letter "ع" ("A'aa" in English) and have solution variances of −2 or larger.

7.2 *Analysis of the Results for Single Users on MPAW and MUEW*

Table 6 shows that words number 3, 4, 5, 8, 9, and 10 (عن, الى, على, ما ان, مع) are recognized without error. In the TAD list, only the words for 4, 5, 8, and 9 contain the letter A'a. The words for the numbers 3 and 8 include the letter (I'ian ع).

Table 6 Words ordered by variance

Word number	Word in Arabic	Optimal solution variance	Recognition rate (%)
4	على	−0.45	100
9	مع	−0.43	100
8	عن	−0.41	100
5	الى	−0.4	100
1	في	−0.39	33
6	التي	−0.36	66
7	الذي	−0.32	66
10	ما	−0.24	100
3	ان	−0.23	100
2	من	−0.14	66

Table 7 Analysis of the recognition rate for the ten most used words in written and spoken English (MUEW)

Word no	English word	Recognition rate (%)	Submission rate (%)	Most often conflicted with	Optimal solution variance
1	The	66	33	Word no. 10	−0.34
2	Of	66	33	Word no. 10	−0.29
3	And	66	66	Word no. 7, 8	−0.33
4	In	66	33 (small error rate)	Word no. 7	−0.30
5	To	66	33	Word no. 7	−0.26
6	Was	100	0		−0.30
7	It	100	0		−0.32
8	Is	100	0		−0.43
9	For	33	66	Word no. 6	−0.07
10	That	100	0		−0.18

Word number 6 التي is pronounced “alatee,” and word number 7 الذي is pronounced “alathee.” Because of their similar pronunciations, these words are often confused with each other. Word number 2 من is the most often confused; its solution variance measure is 0.14 as it can easily be confused with word numbers 1, 5, 6, and 7. The top three words include the letter A, ع, and the following four include the letter ى or ي (measurements of −0.4 to −0.32). The letters ن ا م (pronounced aa, m, and n) yield measurement values between −0.24 and − 0.14. Overall, the average variance in the solution was −0.36 for 90% of the words and − 0.34 for all of them (Table 7).

The words “was,” “it,” “is,” and “that” were recognized without error, but the word “that” had a lower measurement variance (less than −0.08), while the remaining words had variances exceeding −0.25. The variance of the required 90% of the words is approximately −0.31.

In general, we can conclude that Arabic words have a higher recognition rate than English words due to the strong and unique sounds of Arabic characters such as “I’an ع.” Note that Arabic includes more sounds and characters than other languages.

8 Conclusion

In this chapter, a speaker-independent voice recognition system using continuous HMMs has been proposed, developed, and evaluated. After the data are collected, their features are extracted into ten different files, and ten models are trained and tested using maximum likelihood methods, as discussed in Sect. 6. The voice recognition system is trained to maximize the likelihood of a given speech pattern based on the testing environment. The experimental results on multiple users display a correct identification rate of 80% for the list of numbers. This recognition rate can be improved by post-processing, in which more training samples are added. The Arabic language includes a special letter or sound “ع” (“A’aa” in English) shared by other Semitic languages such as Hebrew. We have shown that using words containing the letter “ع” (“A’aa” in English) in a speech recognition system reduces the error rate. On the other hand, words containing soft sounds had lower recognition rates. In this work, relationships were identified between speech recognition and specific sounds in the Arabic language; these relationships will be investigated in further depth for speaker-independent speech recognition in future studies. We have established that the recognition rate depends not only on the tested user but also on the voices used for training.

Acknowledgments This publication was made possible by a grant #NPRP7-442-1-082 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine, 3*(1), 4–16.
2. Bengio, Y., De Mori, R., Flammia, G., & Kompe, R. (1992). Global optimization of a neural network-hidden Markov model hybrid. *IEEE Transactions on Neural Networks, 3*(2), 252–259.
3. Hifny, Y., & Renals, S. (2009). Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(2), 354–365.
4. Holmes, J., & Holmes, W. (2001). Neural networks for speech recognition. In J. Holmes & W. Holmes (Eds.), *Speech synthesis and recognition* (pp. 217–218): CRC Press.
5. Gerhard, D. (2007). *Audio Signal Classification: History and Current Techniques. University of Regina Technical Report TR-CS*

6. Juang, B., Rabiner, L., Levinson, S., & Sondhi, M. (1985). *Recent developments in the application of hidden Markov models to speaker-independent isolated word recognition.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'85.
7. Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall.
8. Juang, B. H., Hou, W., & Lee, C. H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing, 5*(3), 257–265.
9. Al-Ma'adeed, S., Elliman, D., & Higgins, C. A. (2000). A database for Arabic handwritten text recognition research. *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition, 2000, 8*, 130–135.
10. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing, 28*(4), 357–366.
11. Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87*, 1738.
12. Hansen, A. V. (1997). Acoustic parameters optimised for recognition of phonetic features. *Proceedings of Eurospeech-97*, 397–400.
13. Bunke, H., & Wang, P. S. (1997). *Handbook of character recognition and document image analysis*: World Scientific Publishing Company Incorporated.
14. Doddington, G. R. (1998). Sheep, Goats, Lambs and Wolves - An Analysis of Individual Differences in Speaker Recognition Performance, from http://simson.net/ref/1998/Sheep_Goats_Lambs_and_Wolves.pdf
15. Mansour, A. Arabic Script Retrieved 6 June 2005, from http://www.omniglot.com/writing/arabic.htm

Information Visualization Techniques for Building Better Visualization Models

Rachael Fernandez and Noora Fetais

1 Introduction

Recent advances in Information Technology have led to the generation of vast amounts of data that overwhelm the user. This data deluge makes it hard to extract useful information and the user is often swamped with ginormous amounts of data. Much of this data is in textual format, and this makes it harder to make sense of the data at hand.

Card et al. defined information visualization as the “use of computer supported, interactive, visual representations of data to amplify cognition” [1]. Visualization technologies help the users to identify patterns or abnormalities in the data so that the user can make informed decisions. At the high level, Information Visualization (IV) systems are made of two components, namely, Visual Representation and Interaction [2]. Visual Representation refers to the mapping of data to the objects in the scene. However, without any interactivity, the information rendered by the visualization systems are just static and do not fulfill the purpose of communicating the entire picture to the users. For example, a static image graph is much less effective than an interactive graph in which the user can filter the data that is displayed.

This leads to the need for developing interactivity techniques, which help the user in exploring the data. The interaction with the visualized scene starts with the need to explore the data, which helps the user to better understand the data and the visualized scene. Once the user has completed the interaction, the system processes the input and provides feedback to the user. The users often interact with the visualization system to adjust the following options [3]:

R. Fernandez · N. Fetais (✉)
KINDI Center for Computing Research, Doha, Qatar
e-mail: rf1405233@qu.edu.qa; n.almarri@qu.edu.qa

J. M. Alja’am et al. (eds.), *Recent Trends in Computer Applications*,
https://doi.org/10.1007/978-3-319-89914-5_17

1. **Modify the data to be visualized:** The user can explore the various facets of the data by selecting and filtering information based on the needs. This helps in viewing an overview of the data or exploring some subset of the data in detail.
2. **Customize the display:** The users often interact with the system to customize the visualization screen. For example, the screen brightness, background color, or visualization shapes can be changed as per the user's requirements. These modifications help in exploring of the data smoother for the user.

Interactive Visualization systems help to combine representation and interaction of data, where an overview of the data is provided, followed by the details of the selected object as and when they are demanded by the user. The use of interactive options on the visualized object helps us to interact and manipulate the data object as opposed to viewing a static image [3, 4].

A complete Visualization System is constructed by going through a series of stages, which are often referred to as the Visualization Pipeline. According to [8], the pipeline transforms the data (value) into the visualization (view). The Prefuse Toolkit describes a simple visualization pipeline, in which the abstract data that is to be visualized is first filtered and transformed into a visual form. This transformation is done by applying the visual attributes on the objects to be visualized. These objects are finally rendered on the scene and can be viewed by the user [9].

This chapter is structured as follows: Sect. 2 discusses a simple visualization pipeline. A list of attributes to enhance representative and interactive visualization techniques for two-dimensional (2D) scenes is described in Sect. 3. Section 4 introduces visualization techniques for enhancing cognition levels in three-dimensional (3D) scenes. This is followed by a short discussion of the popular visualization tools and the functions that are available in each of these tools in Sect. 5. The chapter concludes with a summary in Sect. 6.

2 Visualization Pipeline

Existing visualization pipelines focus on the representative visualization module of the model and ignore the integration of the interactivity attributes into the model. In this section, we will outline the general steps that are required for creating a visualization model that couple both representative and interactive techniques.

The pipeline goes through the following five steps:

1. **Map to a Geometrical Shape:** The geometrical shape to which the data can be mapped to must first be identified. For example, a polygon with multiple vertices can be used for representing multidimensional data in a fixed-range, whereas concentric-circles can be used for representing different levels of access.
2. **Layout:** If the object is made up of multiple objects, then the layout and the relationship between the different objects have to be defined. For example, in a simple 2D chart, plotting of values along the axes indicate an increase of the value.

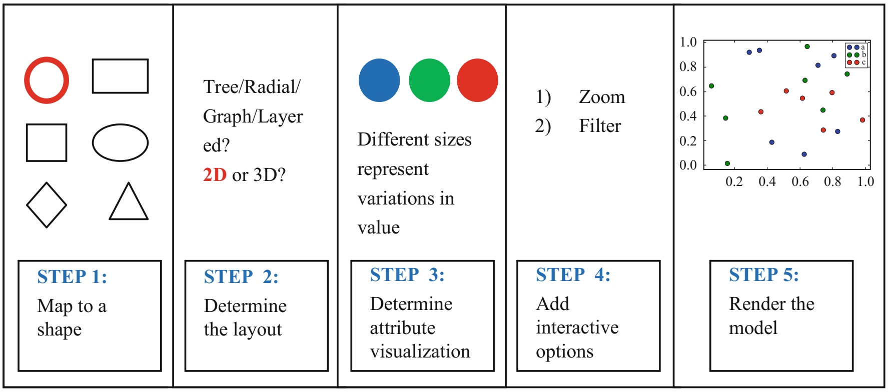


Fig. 1 Steps in the visualization process

3. **Attributes:** It should be ascertained, that all the attributes of the data can be conveyed through the visualized object. For example, the marks of two students can be differentiated by using two circles of different colors and making the size of the circles in proportion to the marks that were obtained.
4. **Addition of Interactive Options:** Interactive options such as zooming, filtering, and panning can be added. Interactive options help in better exploration of the data and increase engagement with the user.
5. **Rendering:** The visualized object is finally rendered on screen after setting parameters like camera perspective, lighting, and depth in the case of multidimensional data.

It must be noted that the pipeline is an iterative process that requires the scene to be updated whenever the user modifies the objects in the scene. Figure 1 depicts the different steps in the visualization process.

3 Visualization Techniques

Visualization techniques are useful for showing an overview of the data, followed by details of the object as and when the user explores the visualized scene. In this section, we will discuss the important attributes required for good representation of the visualized objects and for interaction with these objects in the visualized scene (Fig. 2).

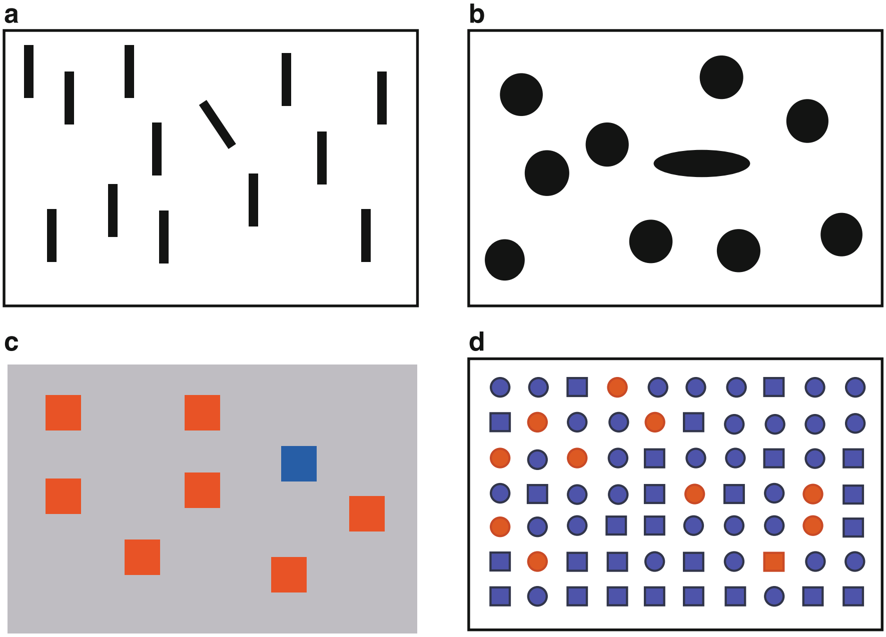


Fig. 2 Importance of using different colors and shapes [5]. (**a**) Different orientations are easier to notice. (**b**) Different shapes are easier to notice. (**c**) Different colors are easier to notice. (**d**) Not as easy to notice the orange square amidst other objects with similar shapes and sizes

3.1 Representative Visualization Techniques

Some of the representative visualization techniques that will enhance the quality of information that is delivered to the user are given in the following [6]:

1. **Shape:** A suitable shape should be used to convey the information. For example, different shapes can be used to signify different objects. For example, the marks of boy students can be represented by a square and that of girl students using a circle in a scatter-plot. Variety of shapes have been experimented with and used to build a visualization model and each of these shapes is better suited for conveying a specific type of information. The iBlogVis system uses a diamond to view the content of a blog and a line to represent the number of characters in a blog [7].
2. **Size**: The size of the objects can be used to indicate the importance or the value of the object in proportion to other objects that are visualized in the scene. The size of the objects refers to the length and width of the object.
3. **Position:** The position of the object in a 2D scene can also be used to represent a parameter. For example, time could be set as the X-axis that would help to find the time at which the event occurred.

4. **Color:** Instead of using different shapes, we can also use different colors (hues) or different intensities of the same color to signify different events or objects.
5. **Orientation:** The orientation of the object in a scene is also easily noticeable.
6. **Depth**: The illusion of depth can be given to objects to visually represent the increasing/decreasing rate of an object.
7. **Texture**: Different textures can be used in different parts of the same object or in different objects to distinguish the objects. It is also used as an indicator of the material of the object. For example, the presence of different types of rocks in a region can be differentiated using textures to mark out the places in a map; it can be used in the medical field for visualizing the anatomical structure [4].
8. **Opacity**: The data that is represented on screen can be represented with different levels of transparency to indicate different levels of saturation of the item. For example, different shades of red can be used for displaying different levels of severity of an event.
9. **Labeling**: The data should be labeled with clear descriptions that will enhance the understanding of the user. For example, the axes should be labeled properly in the case of chart representations.
10. **Layout**: The model is often visualized in different layouts that depend on the type of the data. Some of the popular layouts are **Tree, Radial, and Graph Layouts.** Often, models are created that combine the strengths of these layouts to form a stacked model with different layers [9]:

(a) **Tree Layout:** This layout is best suited for viewing hierarchical data [10, 13] (e.g., Organization Chart, Family Tree). A simple type of tree layout is used for constructing Pedigree Trees that display a family's ancestral roots, as depicted in Fig. 3. It can be modified to add information regarding the birth and death

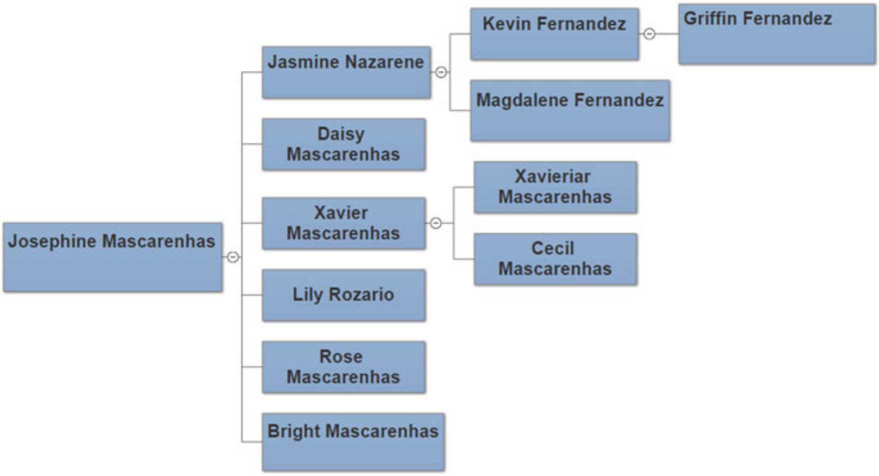


Fig. 3 Pedigree Tree denoting the ancestral roots of a user Josephine

dates of each person. However, these trees expand largely in size as the number of generations increase and are later replaced by Fan Charts, which follow a circular layout that makes them more compact than Pedigree Trees [10].

Fish eye tree views can be used for representing hierarchical textual information. The data in different levels of the hierarchy are depicted by varying degrees of indentations. The data can be viewed in different stages by collapsing or expanding the data in different levels [12].

A variation of trees called Tree-Maps are used for visualizing highly dimensional data (e.g., Medical data) [11, 14]. Tree-Maps are used for visualization of hierarchically *structured* information in a compact rectangular region that is completely filled with information. The region is partitioned into a set of boxes that accommodate nodes. The area occupied by a node in a box depends on the weight or importance of the node.

Tree-Maps are capable of conveying two types of information; *structural information* about the hierarchy and some *content information* associated with each node. Consider the following example that uses a Tree structure (Fig. 4) and nested Tree-Maps (Fig. 5) to visualize a directory structure.

The Tree-Map is adept at visualizing the structure of files inside the various folders in a directory. In a colored Tree-Map, the boxes at the same level in a hierarchy share the same color. In our example, folders B11, B12, B13, C11, and C12 would share the same color (Fig. 5).

While Tree-Maps are 2D structures, Beamtrees on the other hand are 3D in nature. They have the added information of the depth of a node. Beamtrees are created by scaling Tree-Maps.

As Tree-Maps already make use of all the available space in a rectangle, all the boxes inside the rectangle are scaled down in size to a single-dimension structure, which is referred to as the width. After all the boxes have been scaled, all the leaves are assigned to their parent boxes [15].

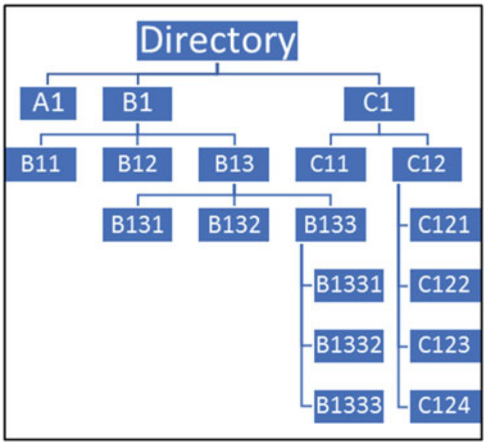


Fig. 4 Tree representation of a file system

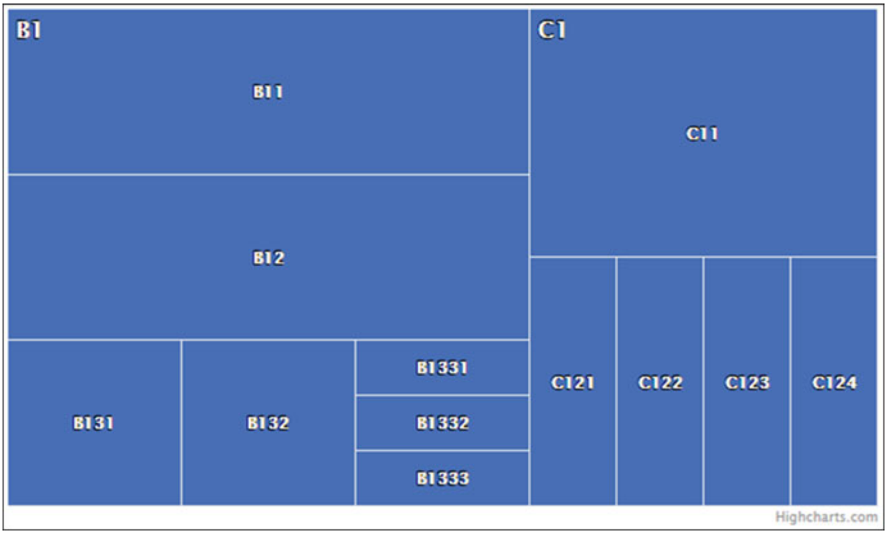


Fig. 5 Tree-Map representation of a file system

Figure 6 depicts a 3D view of the file system in which all the files in the same color are at the same level in the hierarchy.

(b) **Radial Arrangement:** It is a highly versatile model, using which the user can visualize multidimensional data using slices. It is also possible to visualize the different layers of a system using concentric circles. For example, each concentric circle could represent different levels of security [12, 17].

Radial visualization refers to the practice of displaying information within a circular structure. The most popular form of radial displays is the Pie Chart, in which the chart is divided into sectors to visualize the proportions of each element. However, finding the bigger "slice in the pie" is quite difficult. To overcome this difficulty, a pie chart variant that uses concentric circles instead of sectors is used [12].

A combination of these two forms of pie charts is used to visualize events in a file system [17]. In the impromptu system (shown in Figure 7), each slice of the pie is used to represent a single user and the model uses a different color for each user. These slices collapse as a user logs out of the system and the files accessed by each user are visualized as pixels in the slice allocated to the user. Each concentric circle represents a level of permission that is given to the users for accessing a set of files. For example, files in the outermost circle can only be read, whereas the files in the innermost circle can be both read and written on.

Another interesting method is followed in the VisAlert system, which visualizes the network events in a radial enclosure. In this system, an event is described by three attributes; *What*, *When*, and *Where*. Each event is mapped to a 2D space representing the *What* and *Where* attributes as they are finite.

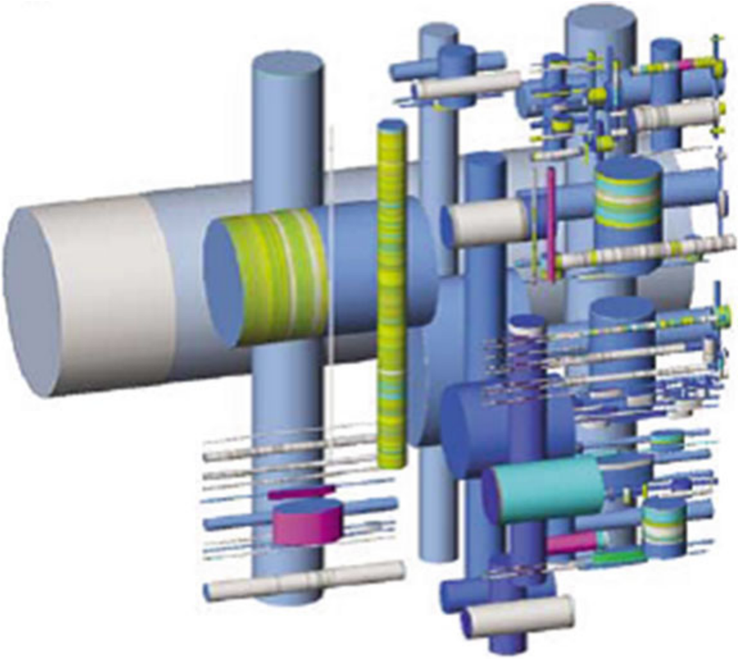


Fig. 6 3D Beamtree representation of a file system using a [15]

A straight line is drawn from this point to the event's *When* attribute to represent the relationship between the three attributes. Figure 8 shows the VisAlert system after the events were visualized [18].

The Visual Thesaurus is an interactive dictionary that displays the selected word in the center and its related words in a radial fashion around the selected word [19]. This helps in specifying the synonyms of the word in a concise manner within a small area. When one of the related words is clicked, the entire arrangement shuffles around to bring the clicked word to the center and its related words around it. Figure 9 represents the Thesaurus structure for the word "information." This approach could be used to visualize related research papers [24].

The Fan Chart (shown in Fig. 10) is used in genealogy to display the ancestors of a person or in simpler words to display the family tree of the person. The outer concentric circle, which is the biggest, contains a crowded list of ancestors. The size of the circle grows with each generation of ancestors. Each ancestor is given the same space as every node in a concentric ring. An interactive variant of the fan charts was proposed by the authors in [20]. In this type of fan charts, clicking on a less important ancestor collapses the central node and removes the node's ancestors from the chart as well. The space freed by the collapsed node is taken over by the neighbor nodes.

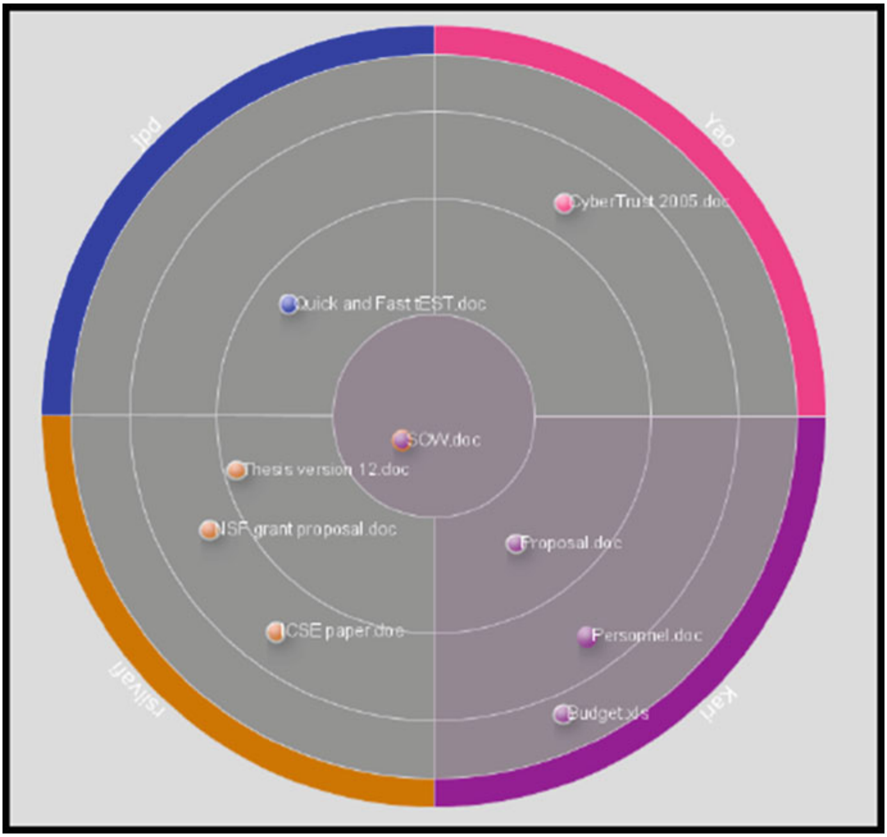


Fig. 7 Visualization interface in the impromptu system [17]

The authors in [21] also used a radial model to detect if a file is a malware by classifying all downloads into good downloads (not malware), bad downloads (malware), and ugly downloads (downloads that we are not sure whether they are good or bad). To analyze if the ugly download is either good or bad, the authors visualize the attributes of the download in a radial fashion. All downloads are plotted in axes inside the radial model, where each axis represents an attribute. The security administrator can then compare the similarities between previous downloads and the current “ugly” download. In Fig. 11, we can see the list of attributes of the download on the left and the radial model with the downloads visualized on the right. The green and red dots in the radial model refer to the good and bad downloads, respectively [21].

(c) **Graph Layout:** Visualization of relationships and interdependencies (e.g., social media networks) and measures of an object or process (e.g., marks, height, accuracy) [11].

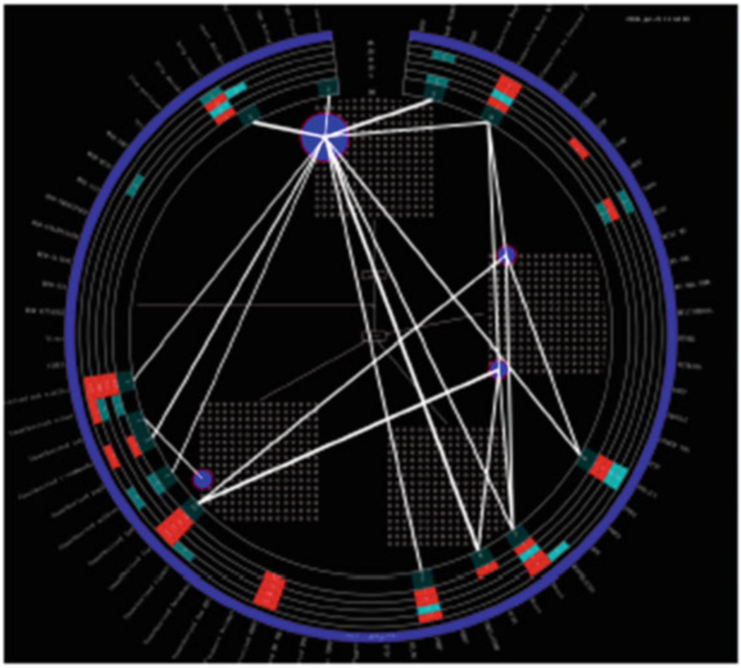


Fig. 8 The VisAlert System [18]

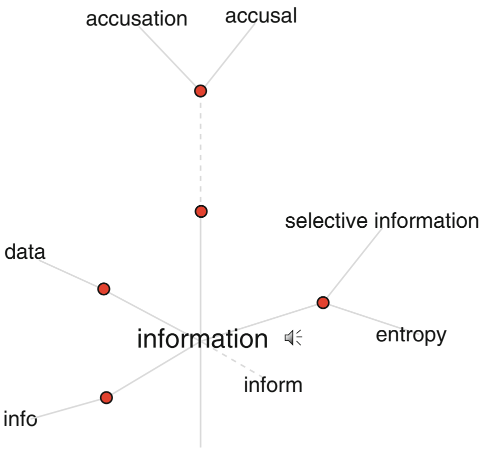


Fig. 9 Visual Thesaurus: Synonyms for the word “information” [19]

Graphs are mostly used to visualize relationships between a set of objects. The objects are represented as nodes and their relationships are represented using edges. A single graph can have multiple types of nodes to represent different sets of objects. Edges are of two types (Fig. 12):

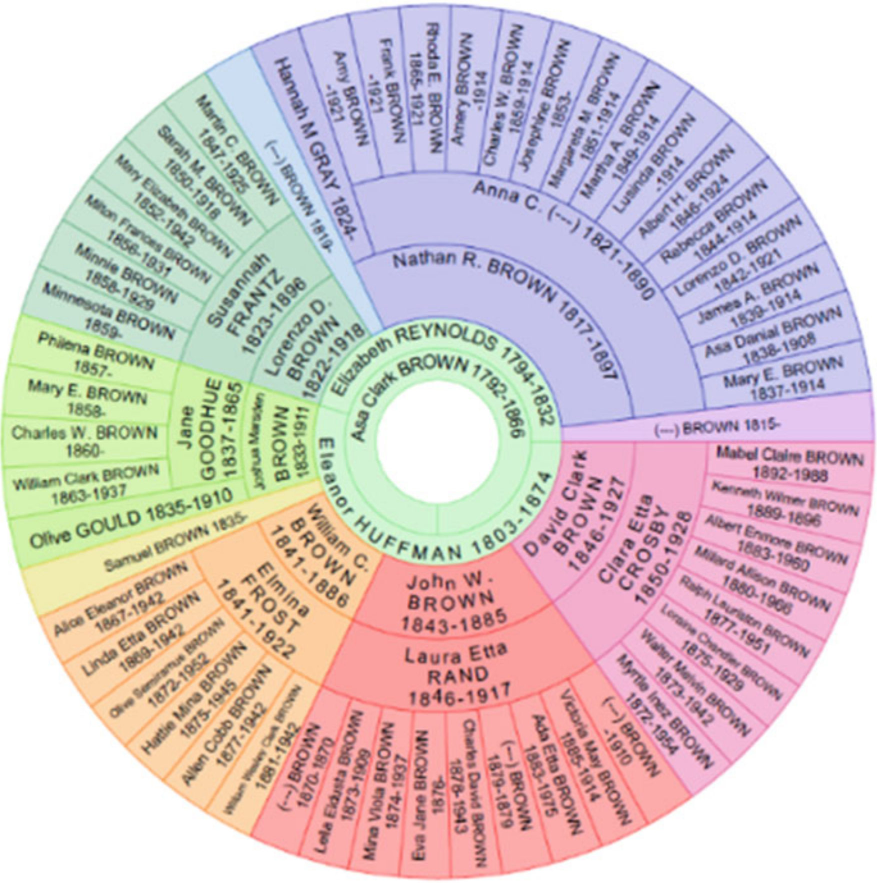
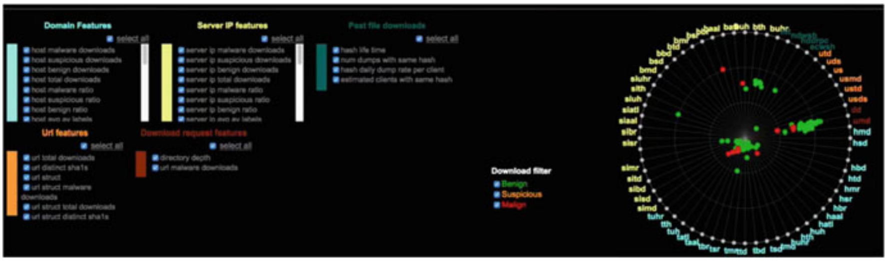


Fig. 10 Fan Chart [20]


Domain Features

☑ select all

☑ host malware downloads
☑ host suspicious downloads
☑ host benign downloads
☑ host total downloads
☑ host malware ratio
☑ host suspicious ratio
☑ host benign ratio
☑ host avg av labels

Server IP features

☑ select all

☑ server ip malware downloads
☑ server ip suspicious downloads
☑ server ip benign downloads
☑ server ip total downloads
☑ server ip malware ratio
☑ server ip suspicious ratio
☑ server ip benign ratio
☑ server ip avg av labels

Past file downloads

☑ select all

☑ hash life time
☑ num dumps with same hash
☑ hash daily dump rate per client
☑ estimated clients with same hash

Url features

☑ select all

☑ url total downloads
☑ url distinct sha1s
☑ url struct
☑ url struct malware downloads
☑ url struct total downloads
☑ url struct distinct sha1s

Download request features

☑ select all

☑ directory depth
☑ url malware downloads

Download filter

☑ Benign
☑ Suspicious
☑ Malign

Fig. 11 Visualization of attributes relating to the downloads [21]

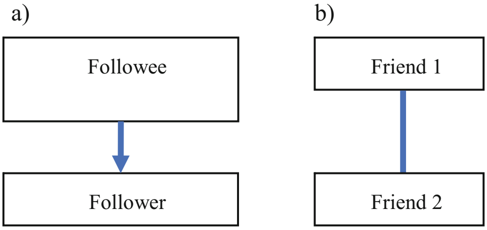


Fig. 12 (**a**) Directed graph; (**b**) Undirected graph

1. **Directed edge:** Indicates a one-way relationship, that is, the edge can be traversed in one direction only. For example, the relationship between followers and the person who is being followed.
2. **Undirected edge:** Indicates a two-way relationship, where the edges can be traversed both ways. For example, these types of edges are used for representing mutual friends in a social media network.

Edges can be manipulated to convey further information. The two common ways edges can be manipulated are:

1. **Weighted edge:** Indicates the strength of the relationship between two nodes. For example, the number of messages exchanged between people in a mutual friend network can be visualized by assigning a weight to it. The higher the number of messages, the thicker is the edge.
2. **Colored edge:** To represent different types of relationships, the edges are colored.

The authors in [22] discuss a bimodal graph (a graph that uses two types of nodes). The graph represents the relationship between community members (represented by circular nodes) and the forums (represented by rectangular nodes) they have posted to. The node size indicates the number of projects that were completed by each member and the weighted edge represents the number of posts that they posted to the forum.

The authors in [23] used a graph to represent the social media network of three Fortune 100s companies to illustrate how these companies are managing their social media platforms—Twitter, Facebook, blogs, and client-hosted forums. Figure 13 indicates how their social media and websites are interlinked.

(d) **Layered Structures:** This layout is good for viewing different layers of an application, levels of security access, etc. This type of visualization is used in the Tudumi system (shown in Fig. 14) [16]. These methods use the visualization techniques described before, but use a set of these methods to form a layered structure to represent more information.

The Tudumi system visualizes user's activities, which are recorded in logs. These events are visualized using a layer of concentric disks, in which the lowest disk represents the user substitution information and the other disks represent other details like the log-in information. The system uses different shapes to identify the different types of users in the system.

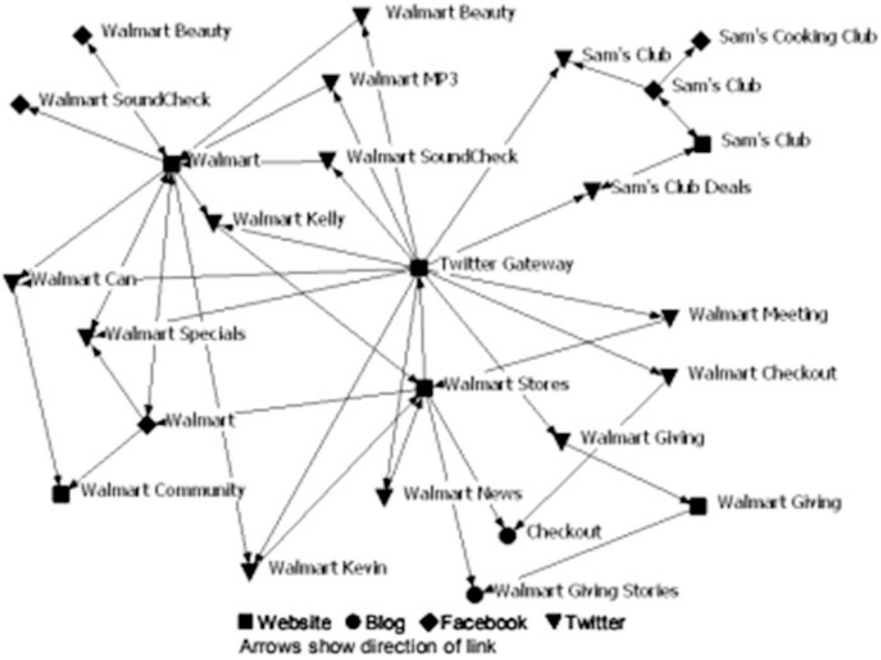


Fig. 13 Graph representing how Walmart's social media presence is linked [23]

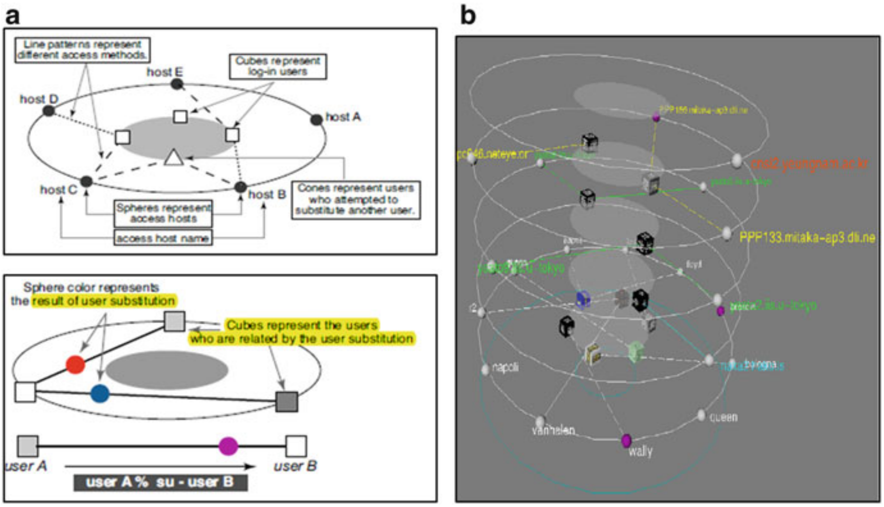

a
Line patterns represent different access methods.
host E
Cubes represent log-in users
host D
host A
host C
host B
Cones represent users who attempted to substitute another user.
Spheres represent access hosts
access host name
Sphere color represents the result of user substitution
Cubes represent the users who are related by the user substitution
user A
user B
user A % su - user B
b
PPP166.mitaka-ap3.dli.ne
pc846.nateye.or
cnsl2.yeungnam.ac.kr
PPP133.mitaka-ap3.dli.ne
napoli
vanhelen
wally
queen
bologna

Fig. 14 The Tudumi system. (**a**) The Visualization paradigm that is followed to visualize events in each layer of security. (**b**) The visualized model that follows a layered structure [16]

A cube represents the logged-in users, spheres represent access hosts, and red cones represent users who attempted to substitute other users. Lines are drawn to represent the relation between access hosts and logged-in users. The color of the cones changes to show successful/unsuccessful substitution.

3.2 Interactive Visualization Techniques

Interactive visualization provides users with more control over the objects in the visualized scene. This helps the user to explore the data and to better understand the results of rigorous experiments. Such techniques not only boost cognition, they also increase user engagement. Users interact with the visualized scene for four main reasons [11]:

1. **Extract more details from an object (Drill-down):** The scene should provide an overview of the data. If the user notices an interesting pattern or object, then the user should be able to drill-down to access more details about the selected pattern or object.
2. **Show less information (Drill-up):** Once the interesting pattern or object has been explored in detail, the user should be able to drill-up to see the overview of the data.
3. **Modify the values of parameters:** The users should be able to modify the default values given to parameters by interacting with the scene.
4. **Customize the scene:** It leads to a much better experience for the user, as interaction is all about the user's preferences. The user should be able to change the color, shape of the objects, etc., in the scene.

Some of the interactive visualization techniques that will enhance the quality of information that is delivered to the user are given below. They are in accordance with Shneiderman's visualization task list [27].

1. **Select:** When there are multiple items on the screen, it may be difficult for the user to keep track of items of potential interest. To overcome this, the user should be able to select an individual item or portions of the data that he or she finds interesting. This enables the user to identify the selected items easily that may be hidden among scores of other items. For example, the user might want to select the outliers in a pattern.
2. **Filter:** The system should allow for filtering data based on the user's choice. It is similar to "select" but the data here is filtered based on a condition as opposed to random selection of events. For example, the user may be interested in only viewing events that occurred during a specified time period [26].
3. **Zooming:** The user should be able to zoom into a pattern or a set of pixels in a dense pixel chart for more clarity. This helps in viewing the items from a low-level area, which provides more information when the patterns are observed from a closer view. The zoomed view can be provided as a pop-up or in the ancillary display to provide extra information regarding the selected pixels [25].

4. **Glossing:** The system should provide some basic details about the component when the cursor hovers on top of it. Some basic information like the time an event occurred, number of components, etc. could be displayed in the tooltip.
5. **Reconfiguration:** In some application, the user might want to focus on some principal event and view its relationship with other secondary events. In this case, the system should support reconfiguration to allow for changing the focus on different events [22]. Consider the example of the Visual Thesaurus, which is an interactive dictionary that displays the selected word in the center and its related words in a radial fashion around the selected word. When one of the related words is clicked, the entire arrangement shuffles around to bring the clicked word to the center and its related words around it. Figure 9 represents the thesaurus structure for the word "information."
6. **Blinking:** The change in the status of any component, in the case of real-time visualization, should be communicated to the user by rapid flickering of the component to grab the attention of the user.
7. **Distortion:** This allows us to focus on a period of time, which provides a zoomed view of the selected period [26].

4 Visualization Options for 3D Scenes

Some of the following options are optional in a 2D setting; however, they are necessary for rendering a 3D scene.

1. **Perspective:** It refers to the angle at which the camera is set with respect to the object. Different perspectives can discover different dimensions of the object. Hence, an appropriate perspective that offers the best insight into the data should be selected.
2. **Panning:** Once the perspective is decided, that is, the camera is fixed at a location, the scene can be viewed from different angles. This helps in getting a better idea of the position and placement of all the objects in the scene by moving across different views [26].
3. **Rotation:** The object can be rotated about any axis to get complete information about the object. In Fig. 15(b), the letter "F" has been rotated by 180° along the y-axis.

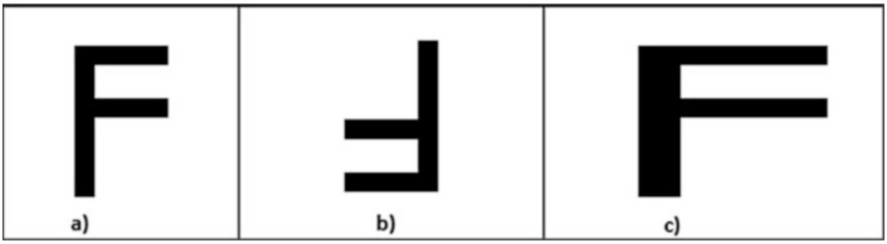


Fig. 15 Effects of rotation and scaling: (**a**) Normal view of letter "F" (**b**) 90° Rotated "F" (**c**) Horizontally scaled "F"

4. **Scaling:** The object can be scaled in size, along any of the axis. In Fig. 15(c) we can see that the letter “F” has been scaled to double its original size along the x-axis.
5. **Lighting:** The objects in 3D are often present in an environment that is “lighted.” This allows the user to view the surface of the object with an illusion of depth.

5 Popular Tools for Creating Visualization Models

In this section, we will look at some of the popular visualization tools that are easy to use and are properly documented. Table 1 shows some of the tools that help in visualizing the data in the form of charts or maps, whereas other tools allow the user to build their own visualization models.

Google Maps and Kartograph help in visualizing geographical data using maps. Tools like Google Charts, Fusion Charts, ZingCharts help in visualizing the data in the form of charts and the user can choose from the various predefined charts. Most of the tools provide highly interactive charts that are easy to work with. ZingCharts and Fusion Charts also help in creating dashboards that help to present the data to the user in a highly concise and compact manner.

Tableau is also used for chart modeling. It is easy to use mainly because of the ease with which you can plug in your data and choose from the different charts that are available. In addition, it provides an extremely easy drag-and-drop feature to select the visualization attributes.

Tools like Processing, D3, and Three.js are used to create their own visualization models in 2D/3D by coding in the supported languages. Some of these tools also provide various features that help in enhancing the interactivity of the created model. Fabric.js is another easy-to-use tool that is used for creating simple visualization models with more support for animation.

Table 1 Popular visualization tools

Tool	Can create new visualization models?	Supports 3D?	Visualization type
Tableau	No	Yes	Charts
Processing	Yes	Yes	User-built model
D3	Yes	Yes	User-built model
Three.js	Yes	Yes	User-built model
Fabric.js	Yes	No	User-built model
Google Charts	No	Yes	Charts
ZingCharts	No	Yes	Charts
Kartograph	No	Yes	Maps

6 Conclusion

As data is growing at a rapid pace, it is important to visualize the data in a manner that is easy to use, navigate, and understand. It also lays the need for tracking multiple attributes of the data at every second, which leads to the need for creating dashboards that present concise and compact displays of the data. In this chapter, we presented a list of interactive and representative visualization techniques that enhance the quality of the visualization model that can be created by the user. We have also provided a simple visualization pipeline that outlines the steps to be followed to create a model that couples representative and interactive techniques. IV techniques that are specific to 3D visualization models were also presented.

We have also created a list of the popular visualization tools that are available today. A short comparison between some of these popular tools was also presented to help the user in selecting a tool that best suits his or her interests.

Acknowledgments This work was made possible by QUCP-CENG-CSE-15/16-2 grant from the Qatar University Fund. The statements made herein are solely the responsibility of the authors.

References

1. Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.
2. Chegini, M., Shao, L., Lehmann, D. J., Andrews, K., & Schreck, T. Interaction Concepts for Collaborative Visual Analysis of Scatterplots on Large Vertically-Mounted High-Resolution Multi-Touch Displays.
3. Zudilova-Seinstra, Elena, Tony Adriaansen, and Robert Van Liere. “Overview of Interactive Visualisation.” Trends in Interactive Visualization. Springer London, 2009. 3–15.
4. Brodbeck, Dominique, Riccardo Mazza, and Denis Lalanne. “Interactive visualization-A survey.” Human machine interaction. Springer Berlin Heidelberg, 2009. 27–46.
5. Hicks, M. (2009). Perceptual and design principles for effective interactive visualisations. In Trends in Interactive Visualization (pp. 155–174). Springer London.
6. Pham, B., Streit, A., & Brown, R. (2009). Visualisation of Information Uncertainty: Progress and Challenges. In Trends in interactive visualization (pp. 19–48). Springer London.
7. Vassileva, Julita, and Carl Gutwin. “Exploring blog archives with interactive visualization.” Proceedings of the working conference on Advanced visual interfaces. ACM, 2008.
8. Chi, E. H. H., & Riedl, J. T. (1998, October). An operator interaction framework for visualization systems. In Information Visualization, 1998. Proceedings. IEEE Symposium on (pp. 63–70). IEEE.
9. Heer, Jeffrey, Stuart K. Card, and James A. Landay. “Prefuse: a toolkit for interactive information visualization.” Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2005.
10. Tuttle, C., Nonato, L. G., & Silva, C. (2010). PedVis: a structured, space-efficient technique for pedigree visualization. IEEE transactions on visualization and computer graphics, 16(6), 1063–1072.
11. Keim, Daniel A. “Information visualization and visual data mining.” IEEE transactions on Visualization and Computer Graphics 8.1 (2002): 1–8.

12. Wu, Yingcai, et al. “OpinionSeer: interactive visualization of hotel customer feedback.” IEEE transactions on visualization and computer graphics 16.6 (2010): 1109–1118.
13. Nguyen, Quang Vinh, and Mao Lin Huang. “EncCon: an approach to constructing interactive visualization of large hierarchical data.” Information Visualization 4.1 (2005): 1–21.
14. Fernandez, Rachael, and Noora Fetais. “Framework for Visualizing Browsing Patterns Captured in Computer Logs Using Data Mining Techniques.” International Journal of Computing & Information Sciences 12.1 (2016): 83.
15. Van Ham, F., & van Wijk, J. J. (2003). Beamtrees: Compact visualization of large hierarchies. Information Visualization, 2(1), 31–39.
16. Takada, Tetsuji, and Hideki Koike. “Tudumi: Information visualization system for monitoring and auditing computer logs.” Information Visualisation, 2002. Proceedings. Sixth International Conference on. IEEE, 2002.
17. De Paula, Rogerio, et al. “In the eye of the beholder: a visualization-based approach to information system security.” International Journal of Human-Computer Studies 63.1 (2005): 5–24.
18. Livnat, Y., Agutter, J., Moon, S., Erbacher, R. F., & Foresti, S. (2005, June). A visualization paradigm for network intrusion detection. In Information Assurance Workshop, 2005. IAW’05. Proceedings from the Sixth Annual IEEE SMC (pp. 92–99). IEEE.
19. Thinkmap Inc., “Thinkmap Visual Thesaurus,” http://www.visualthesaurus.com
20. Draper, G. M., & Riesenfeld, R. F. (2008). Interactive fan charts: A space-saving technique for genealogical graph exploration. In Proceedings of the 8th Annual Workshop on Technology for Family History and Genealogical Research (FHTW 2008).
21. Angelini, M., Aniello, L., Lenti, S., Santucci, G., & Ucci, D. (2017, October). The goods, the bads and the uglies: Supporting decisions in malware detection through visual analytics. In Visualization for Cyber Security (VizSec), 2017 IEEE Symposium on (pp. 1–8). IEEE.
22. Hansen, D. L., Rotman, D., Bonsignore, E., Milic-Frayling, N., Rodrigues, E. M., Smith, M., & Shneiderman, B. (2012, December). Do You Know the Way to SNA?: A process model for analyzing and visualizing social media network data. In Social Informatics (SocialInformatics), 2012 International Conference on (pp. 304–313). IEEE.
23. Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. MIS Quarterly Executive, 9(4).
24. Yi, Ji Soo, Youn ah Kang, and John Stasko. “Toward a deeper understanding of the role of interaction in information visualization.” IEEE transactions on visualization and computer graphics 13.6 (2007): 1224–1231.
25. Carlis, John V., and Joseph A. Konstan. “Interactive visualization of serial periodic data.” Proceedings of the 11th annual ACM symposium on User interface software and technology. ACM, 1998.
26. Bade, Ragnar, Stefan Schlechtweg, and Silvia Miksch. “Connecting time-oriented data and information to a coherent interactive visualization.” Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2004.
27. Shneiderman, Ben. “The eyes have it: A task by data type taxonomy for information visualizations.” Visual Languages, 1996. Proceedings., IEEE Symposium on. IEEE, 1996.