ProGes:
A User Interface for Multimedia Devices over The Internet of Things

by

Ali Ahmadi Danesh Ashtiani

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the M.Sc. degree in
Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

With the rapid growth of online devices, a new concept of Internet of Things (IoT) is emerging in which everyday devices will be connected to the Internet. As the number of devices in IoT is increasing, so is the complexity of the interactions between user and devices. There is a need to design intelligent user interfaces that could assist users in interactions. Many studies have been conducted on different interaction techniques such as proxemic and gesture interaction in order to propose an intuitive and intelligent system for controlling multimedia devices over the IoT, though most could not propose a universal solution. The present study proposes a proximity-based and gesture-enabled user interface for multimedia devices over IoT. The proposed method employs a cloud-based decision engine to support user to choose and interact with the most appropriate device, reliving the user from the burden of enumerating available devices manually. The decision engine observes the multimedia content and device properties, learns user preferences adaptively, and automatically recommends the most appropriate device to interact. In addition to that, the proposed system uses proximity information to find the user among people and provides her/him gesture control services. Furthermore, a new hand gesture vocabulary is proposed for controlling multimedia devices through conducting a multiphase elicitation study. The main advantage of this vocabulary is that it can be used for all multimedia devices. Both device recommendation system and gesture vocabulary are evaluated. The device recommendation system evaluation shows that the users agree with the proposed interaction 70% of the times. Moreover, the average agreement score of the proposed gesture vocabulary (0.56) exceeds the score of similar studies. An external user evaluation study shows that the average score of being a good-match is 4.08 out of 5 and the average of ease-of-performance equals to 4.21 out of 5. The memory test reveals that the proposed vocabulary is easy to remember since participants could remember and perform gestures in 3.13 seconds on average. In addition to that, the average accuracy of remembering gestures equals to 91.54%.

Acknowledgements

This thesis would not have been possible without the help of many people to whom I am greatly indebted. I would like to express deepest appreciation to my supervisor Professor Abdulmotaleb El Saddik for his continuous support during my graduate studies and research. I appreciate his patience, motivation, enthusiasm, and immense knowledge. His guidance has helped me tremendously in the research and writing of this thesis.

I would like to express my infinite appreciation and love to my parents, Mostafa and Nahid, whose consistent encouragement, support, and unconditional love have assisted me with the hurdles encountered during my academic journey. Also, I would like to extend my gratitude to my dearest siblings, Zahra and Mahdi, and my great brother-in-law Ahmadreza for their continues support during these years.

I would like to express my appreciation to Dr. Mukesh Saini, who collaborated closely with me on the better part of my research. He has continuously put his solid technical qualifications at my disposal throughout my academic endeavor. Also, I would like to thank Dr. Hussein Al Osman, Dr. Benjamin Guthier, and Dr. Haiwei Dong for their kind support and help. Moreover, I would like to thank all the members of the Multimedia Communications Research Laboratory for their cooperation and support and for simply being wonderful friends.

Last but not least, I would like to thank my supportive friends Hossein Bonakdar, Hadi Rahmatkhah, Javad Sheikhtaheri, Pedram Roshani, Morvarid Kardan, Valeh Montaghami, Melissa Dawe, Murat Varol, and Mohammad Huda.

Table of Contents

List of Tables

List of Figures

Nomenclature

Abbreviations

2D	2 Dimensional
3D	3 Dimensional
3G	3rd Generation
AP	Access Point
AR	Augmented Reality
DMS	Degree Decimal Minutes
DSL	Digital Subscriber Line
DTW	Dynamic Time Warping
GCU	Gesture Control Unit
GIU	Gesture Interaction Unit
GUI	Graphical User Interface
HCI	Human Computer Interaction
HD	High Definition
HMM	Hidden Markov Model
ID	Identification
IoT	Internet of Things
IPU	Interaction Proxy Unit
IR	Infera Red
IT	Information Technology
LTE	Long-Term Evolution
MRU	Media Redirection Unit
NUI	Natural User Interface
PIU	Proxemic Interaction Unit
RF	Radio Frequency
RFID	Radio Frequency Identification
RGB	Red Green Blue
RSS	Received SIgnial Strength
RSSI	Received SIgnial Strength Indicator
SNR	Signal to Noise Ratio
TOF	Time Of Flight
UCM	Use Case Map
UF	Ultrasound Frequency

UI	User Interface
UTM	Universal Transcerse Mercator
WiFi	Wireless Fidelity
WiMax	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

Mathematical Symbols

a_{iu}	user's age coefficient
A_r	agreement score of gesture r
C	the set of available multimedia contents
C_c^a	multimedia content has audio
c	the multimedia content that user is playing
C_c^d	multimedia content length
C_c^v	multimedia content has video
D	the set of available devices
D_i	device i
D_i^a	device i can play audio
d_{iu}	distance effectiveness coefficient
D_i^v	device i can play video
f_{ic}	device playback capability flag
h_{iu}	user's multimedia usage history coefficient
m_{ic}	device appropriateness coefficient
P_i	subset of identical gestures within P_r
p_{iu}	user's profession coefficient
P_r	set of proposed gestures for the referent r
s_i	score for device i
u	the user
x_{iu}	distance between device i and user u

Chapter 1

Introduction

The complexity of the interactions between user and devices is increasing as the number of online devices rises. Hence, there is a need to design intelligent user interfaces that could assist users in interactions. This thesis presents a proximity-based and gesture-enabled user interface for multimedia devices over the Internet of Things. This chapter presents a brief overview on this study. It starts with a short introduction on the Internet of Things. Then, it explains our motivation for conducting this study followed by the problem statement. The problem statement is divided into two smaller parts to make it perfectly understandable. Next, the potential contributions of our study are presented. In the end, the thesis outline is given.

1.1 Internet of Things

The Internet was invented back in the 1960s but its usage was very restricted [22]. The first limitation was the quantity of available resources, contents, and services. The other major restriction was the number of people who had access to it. However, it has been expanding quickly since then and recent advances in technology have helped it grow quite a bit. In modern times, people can access to the Internet through various devises (e.g. cellphones, tablets, laptops, smart TVs, etc.) and technologies (e.g. DSL, WiMAX, 3G, LTE). As a result, there are more than 2.7 billion regular Internet users all across the globe, which is equal to 39% of worlds population [56].

Moreover, advances in technology have increased the size of the Internet by providing more available resources, contents, services and online devices. Since the number of online objects (e.g. devices, services, contents, etc.) has been growing dramatically, a new concept was introduced in 1990s: the Internet of Things (IoT) [42]. In 2013, there were more than 11.2 billion devices connected to the Internet and it is predicted that there will be around 50 billion devices online in 2020, all of which will be part of the IoT [46].

1.2 Motivation

More than 7 billion people are living all around the world and they constantly interact with each other and their surrounding environment using both verbal and nonverbal communication channels. Verbal communication channels like speaking and writing are considered the main channel since they are usually used explicitly. Nonverbal communication, however, is more often used unconsciously and implicitly. As a result, we must look at interactions with our environment with greater focus in order to notice and understand nonverbal communication interaction. Haptic, Proxemic, Body Language, etc. are some of the examples of our nonverbal communication channels. They can be used solely to communicate or can be a great supplementary to verbal communication by providing valuable information. Thus, they play a significant role in people's lives.

The number of online objects is increasing quickly and they are developing further "intelligent" competences. As a result, the IoT is becoming more and more similar to human society. However, comparison of these two societies shows that nonverbal communication is not involved in interactions over the IoT. Yet, it can be argued that the IoT can benefit from a greater presence of nonverbal communication tools since they would help give the currently existing dialogues between individuals and groups a more natural and realistic structure. Therefore, developing nonverbal communication channels within the IoT is a top priority for many researchers, including ourselves.

1.3 Problem Statement

As we explained in section 1.2, the size of the IoT is expanding very quickly. The uniquely identifiable objects and devices move and interact with other things within the IoT. In the other word, the IoT is like a big and dynamic society of things. Yet it is far from the ubiquitous computing vision of Weiser [71] due to two main reasons: the lack of an appropriate task-centered user interface design approach and the lack of reliable support for distributable user interfaces in ubiquitous environments [39]. Thus, there is a need for a new generation of specifically designed user interfaces for IoT.

To propose a new user interface, we should consider the primary priority of the Human Computer Interaction (HCI): User's point of view. Users have different attitudes and diverse preferences, so those attitudes and preferences should be studied and elicited very well. The proposed solution should provide a Natural User Interface (NUI) to satisfy the users. Moreover, since we want to propose a UI for multimedia devices over the IoT, it should be distributed and embedded in multimedia devices to provide a homogeneous environment for users. Then, a main intelligent decision maker engine can assist users to make use of their surrounding devices better by providing them the most appropriate device, reliving the user from the burden of enumerating available devices manually. These two problems are discussed in details in the following subsections.

1.3.1 Natural User Interface

Our daily lifestyle has been influenced by recent breakthroughs in technology. New interaction devices and technologies such as Microsoft Kinect sensor, Apple's Siri speech recognition engine, etc. have assisted researchers to develop new natural user interfaces. However, researchers are still trying to find an intuitive engagement mechanism for these new interaction devices. Therefore, we should find a solution for engagement mechanism of our system, as we want to emerge these natural interaction devices in our system. We decided to use proxemics as the engagement mechanism in our system.

Moreover, since speech recognition uses language's standards, there are a lot of devices and applications that are speech enabled. Gesture enabled devices, however, have a completely different story. So far, they rely on hand tracking or identifying a minimal set of hand-gestures, all of which are device specific and yet not standardized. As we will review in the next chapter, several studies have proposed gesture vocabularies for controlling multimedia and entertainment devices but each of them has its drawbacks (e.g. size, performance complexity, etc.). As a result, in order to propose and develop a new NUI for multimedia devices over the IoT, an interactive hand-gesture vocabulary should be derived and implemented.

The proposed gesture vocabulary, G, should have some characteristics and features. First, it should be practical. In the other words, it should be able to cover all of the basic functionalities and commands. Second, it should be intuitive which means that each command and its corresponding gesture should have a meaningful connection or relation. Third, it should be easy to perform, so performing them should not exhaust the users. Last, it should be easy to remember for the users. Thus, users could remember and perform them without spending much time on thinking. There is a straight relation between the number of supported commands and the size of gesture vocabulary. When, the size of vocabulary increases, the supported commands can be expanded as well. However, defining more gestures means that performance complexity increases and it is harder to remember all of them. Hence, there is a trade-off between the size of gesture vocabulary and its usefulness. As a result, the proposed vocabulary should be intuitive and have a proper size to increase the memorability while it holds its functionality. Additionally, it should be simple and easy to perform to be seamlessly adopted by the users, making it to be universally standardized. Figure 1.1 shows the problem at the high-level domain.

1.3.2 Intelligent Decision Maker

As we explained earlier, since the proposed UI should work over the IoT, it should be a distributed and embedded UI within the multiple devices that can receive and interact with the multimedia stream. Moreover, users have different preferences and attitudes towards multimedia devices. As a result, there should be a personalized multimedia device recommender system. Using proxemic interaction, we can track the location of different devices, and then find and recommend an alternative device for the media stream if any exists. Let $D = \{D_i | 1 \leq i \leq n\}$ be the set of all available devices, u be the user, and c

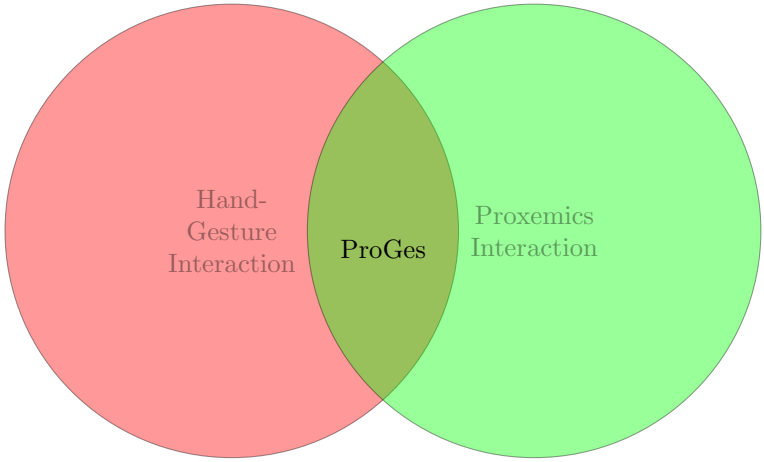


Figure 1.1: The proposed UI, *ProGes*, is the intersection of two different interaction techniques: Hand-Gesture interaction and proxemic interaction.

be the multimedia content that user is playing. Then, we want to find a device scoring function as follows:

$$s_i \;=\; f(D_i, u, c) \tag{1.1}$$

where s_i is the score of D_i for user u and content c. This score will be used to find the best available device and suggest that device to the user. The device which is recommended by the proposed system should be the same as the preferred device of the user. However, it is impossible to consider all users in the initial state of our system. Hence, our proposed system should be able to adaptively enhance itself over the time in order to learn user's preferences and recommend the preferred device of the user.

1.4 Contributions

This thesis introduces a proxemic and gesture enabled user interface (ProGes) for multimedia entities within the Internet of Things. The main goal of this research is to elicit users' attitudes and preferences, and then propose a user interface. The proposed UI is designed to control multimedia devices and applications using a combination of proxemic and gesture interactions. To achieve this goal, we have:

- Studied multimedia devices and application and their varying requirements.
- Conducted two multiphase user studies to elicit user attitudes and preferences towards multimedia devices and hand gestures.
- Designed and implemented a new algorithm for engaging users with devices.

- Designed and implemented an adaptive device recommender system that works based on two algorithms: device scoring algorithm and an adaptation algorithm.
- Designed and implemented a gesture vocabulary for controlling multimedia playback devices.
- Designed and implemented ProGes on two platforms: Mac OS X and Android.
- Conducted two external user studies to evaluate ProGes and validate its goals.

1.5 Scholarly Achievements

In the process of completing this study, the following publication has been submitted and accepted:

- **A. Danesh**, M. Saini, A. El Saddik, "A Proxemic Multimedia Interaction over the Internet of Things," Multimedia Modeling (MMM), 21^{st} Anniversary International Conference on, 5–7 January 2015, to be published.

1.6 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 has two sections: the first section focuses on proxemic interaction. It begins by defining proxemic interaction and its two types of applications. Next, it studies related works and, in the end, highlights the differences between those related works and the present study. The second section of this chapter is dedicated to gesture elicitation studies. It starts with an introduction on gesture elicitation studies and their importance. Then, it presents an overview on a number of pioneering and similar studies. This section is summarized by a comparison between the reviewed works and the present study to make their differences clear.

Chapter 3 presents the details of conducted user studies. In the first section, it explains the details of an elicitation study on user preferences and attitudes towards a gesture vocabulary for multimedia devices. This section ends by introducing our gesture vocabulary. A user study on multimedia device preferences is presented in the second section of this chapter. We conducted a broad user survey to find some of the elements most likely to influence users' decisions when selecting multimedia devices. At the end of this section, the results and findings of this user study are given.

Chapter 4 presents the proposed solution, ProGes. First, the abstract architecture of the proposed system is presented. Then, the functionality description of the system is given with the help of two use-case scenarios. Next, different units of ProGes are explained in detail in the following order: Interaction Proxy Unit, Proxemic Interaction Unit, Media Redirection Unit and Gesture Control Unit.

Chapter 5 is dedicated to evaluation and discussion. It starts with a two-step evaluation on gesture vocabulary. First, an external user agreement is presented to validate the proposed vocabulary. Then, it gives the description and results of a memory test on the gesture vocabulary. At the end of this section, some of the results, findings, and challenges of the designing of the gesture vocabulary are given. The second section of this chapter presents an evaluation on device recommendation system and scoring mechanism which is followed by a short discussion on the results and challenges of multimedia device recommendation system.

Chapter 6 gives a brief conclusion of this study based on three perspectives: user engagement mechanisms, device recommendation systems, and gesture vocabulary. This chapter ends with a short discussion about possible future work in ProGes.

Chapter 2

Literature Review

This chapter presents a literature review on the interaction domains that are used in the proposed UI (ProGes) and gives a comparison between the similar studies and this study. The proposed UI is a multifaceted interaction interface (see Figure 1.1). To the best of my knowledge, there exist no other UI that combines proxemic and gesture interaction frameworks together to propose a new UI for multimedia devices over the IoT. However, some studies have used one or the other of these frameworks to control multimedia devices. Consequentially, we have divided this review into two parts (based on the interaction medium) and present them in separate sections. The main goal of this chapter is to clarify the differences between this study and pre-existing related ones. In order to fulfill this goal, we present a summary at the end of each section of this chapter which highlights the main differences and contributions of the present study.

2.1 Proxemic Interaction

Spatial relationship interaction (i.e. proxemic interaction) is widely used in our everyday life (e.g. automatic light switches). As a result, researchers have been trying to increase the involvement of this type of interaction in HCI mechanisms. The architecture of the proposed UI is based on proxemic interactions. Unfortunately, it is not possible to make the proxemic interaction between user and devices unless their location information is also available. We have also used various positioning techniques in the prototype system. They are used as proofs of concept to validate ProGes. Thus, in this section, the background of positioning systems and techniques is presented, after which we start the study of related works in proxemic interaction.

2.1.1 Background

Finding the current location of a user and keeping it updated is a very classic yet controversial problem. User location information is advantageous since it can refine and personalize services that are provided to users but on the flip side, it can also violate users' privacy.

Hence, researchers are always debating whether or not user location tracking is ethical. These arguments, however, did not stop researchers from working on this problem. Academic and industry research centers applied different techniques and technologies to find the user location, which resulted in several methods that can provide the location information for location-based applications with the required level of accuracy. There are different types of location information: physical, symbolic (proximity), absolute, and relative [24].

The physical location information is represented by a 2D/3D point coordinate on a map, which can be measured in either Degree Decimal Minutes (DMS) or Universal Transverse Mercator (UTM) format. Symbolic or proximity location information is expressed in natural language words such as kitchen, living room, etc. Absolute location is measured by a shared reference grid in the locating area and, finally, relative location information calculations are based on local references. [38].

Positioning systems use different techniques to estimate location information. There are three principal categories of techniques for automated locating systems: triangulation, proximity, and vision or scene analysis [23]. Figure 2.1 presents a detailed taxonomy of positioning systems, which are categorized based on the techniques used.

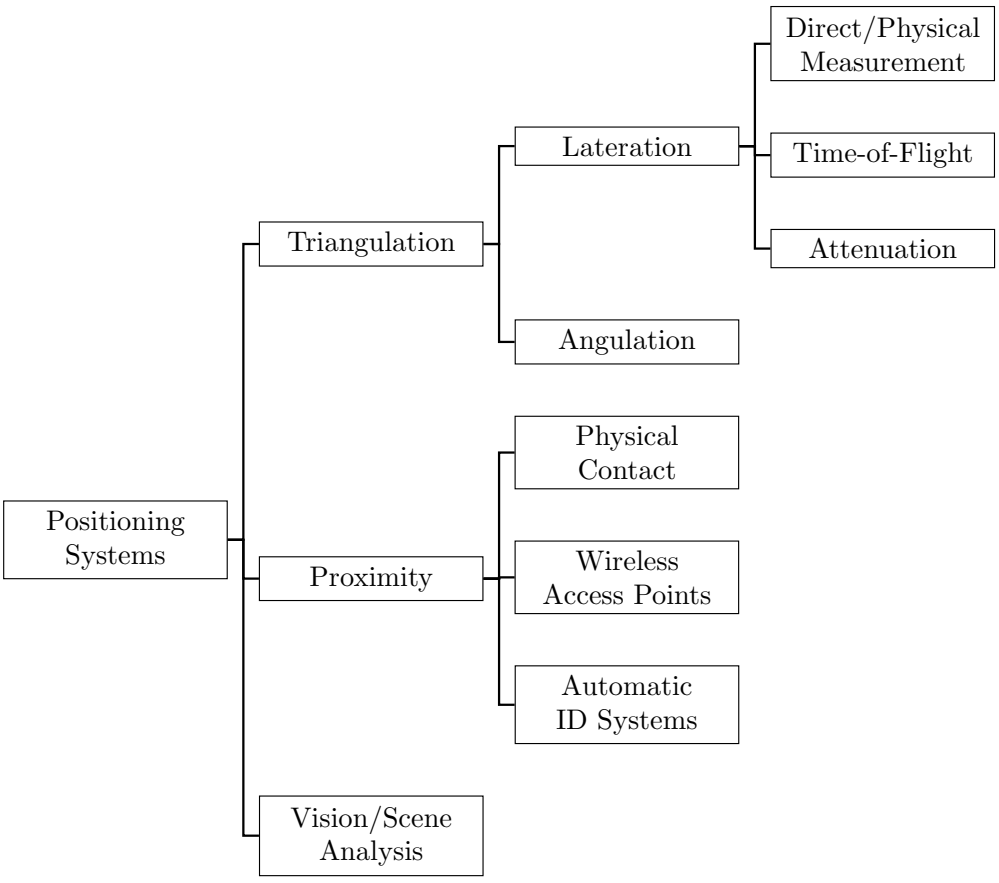


Figure 2.1: The taxonomy of positioning systems based on their estimation techniques.

Triangulation techniques use the geometric properties of triangles to calculate the location information of the object. Triangulation is divided into two subcategories: Lateration and Angulation. There are a lot of similarities between these two categories but there is a slight difference between them. Lateration measures the distance of the object from multiple reference positions while angulation works with the angle or bearing measurements to calculate the location information of the object [23]. Lateration generally uses three different approaches in order to measure the distance between the object and references:

- **Direct Measurement:** A physical action or movement measures the distances. For example, a robot moves and extends its arm to touch something. Then, it calculates its distance from that object. This approach is simple to understand but very complex to implement.
- **Time-of-Flight (TOF):** Measuring the time it takes to travel between the object and a reference point, P, at a known velocity can be used to compute the distance between two locations. It may measure the time for one-way travel or return travel. While agreeing on the correct clock time, blocking and reflection are known problems of this approach, which remains nonetheless very popular.
- **Attenuation:** This approach measures the intensity of emitted signal at the destination point. Then, it measures the distance between the object and the reference point by using the level of the decrease in intensity of the emitted signal. It must be noted that environmental obstructions can easily reduce the accuracy of this approach.

The second group of positioning systems are proximity location sensing techniques. In these approaches, the location is determined by finding the nearest known reference object to the concerned target object. Proximity location sensing is generally divided into three subcategories:

- **Detecting Physical Contact:** Direct physical contact information can be used to detect proximity. Pressure sensors, touch sensors, etc. are some examples of the tools available to this method.
- **Monitoring Wireless Access Points:** There are two types of implementation for this method of which fingerprinting is the first. It consists of two phases: Offline Training and Online Location Information Determination. A set of practical location related data such as access point signal strength, etc. is collected, measured and imported into the system during the offline training phase. Then, at the online positioning phase, a classifier algorithm finds the nearest location to the object using its knowledge of the training phase data. In the second type of implementation, the location information is estimated by monitoring the availability of different access points and their signal strength, which eliminates the need for an offline training phase [19].
- **Observing Automatic ID Systems:** The third method uses automatic identification systems such as key cards, credit cards, etc. to find the closest reference and calculate the location information.

Scene Analysis location sensing (also called vision analysis) is the third technique frequently used by positioning systems. This approach makes use of a set of features extracted from the observed scene. Next, those features are used to draw a conclusion about the position of the observer and objects in the scene. The vantage point is very important in this approach and it can affect the results dramatically. The main advantage of this approach is that it can provide very precious extra information about the location and environment of the target object (e.g. surrounding objects and their relative positions, etc.) when compared to other methods. However, the complexity of the implementation and feature database generation required by this approach prove to be a great disadvantages [23].

Regarding the development of a positioning system, it is important to first decide about its main medium. There are many different media available for positioning systems such as IR frequency, Ultrasound Frequency (UF), and Radio Frequency (RF). Consequently, any of the aforementioned techniques could be used as they all support the required accuracy. Since we wanted to minimize the cost of the system, we decided to use radio frequency as it is the cheapest of the three aforementioned resources. There are three famous subdomains under the RF: Radio Frequency Identification (RFID), Bluetooth, and Wireless Local Area Network (WLAN). We decided to use WLAN since the first two media required either extra accessories or infrastructure extension while, based on primary our assumption (i.e. working over the IoT where all of the objects are connected to the network), WLAN did not need any.

There are numerous positioning algorithms over WLAN. They use different properties and characteristics of WLANs as inputs (e.g. Signal-to-Noise Ratio (SNR), Received Signal Strength (RSS), etc.). Then, they use different calculation formulas or classifiers to estimate the location information. COMPASS [29], Ekahau [1], Horus [79, 78], NearMe [32], and RADAR [4] are just a few examples of these algorithms, all of which have different requirements and diverse accuracies.

In conclusion, since we needed to implement a positioning system that works without any preliminary training in real-time and does not need any configuration, we decided to use Microsoft's NearMe [32] algorithm to design and validate our proxemic interaction system.

2.1.2 Proxemic Interaction Studies

It was 1966 when Hall highlighted the influence of proxemic behavior on interpersonal communication [21]. In his opinion, there exist two valuable aspects of knowledge justifying the study of proxemic interaction. The first one is the ability to evaluate the way people interact with each other in their daily life (micro-level). The second one is studying the space organization of houses, buildings, and ultimately towns (macro-level).

Greenberg et al. proposed a practical version of proxemic interaction for ubiquitous environments that can catalogue people, digital devices and non-digital objects [18]. In their proposal, proxemics is broken down into five dimensions in this proposal: Distance, Orientation, Movement, Identity and Location. Any measured changes in any of these

dimensions can influence the inaction interaction. Using this terminology, Marquardt et al. developed a proximity toolkit that helps ensure the rapid prototyping of proxemic interactions [40]. Then, they went another step forward by using sociological constructs, F-formations and micro-mobility in order to design a tool offering better cross-device interaction [41]. Although recently researchers have been trying to present a more precise structure and organization for proxemic interaction, proxemics have been involved in applications for quite a while at both the micro and macro levels.

Proxemics applications offer many tools for analysis. One such tool is Range [26]. Ju et al. designed Range as a public interactive whiteboard that supports co-located, ad-hoc meetings. It uses the proximity sensing to proactively manage the transitions between display and authors (e.g. to clear the space for writing). The next one tool is Medusa which is a proximity-aware multi-touch tabletop [3]. Ackad et al. proposed another tabletop proxemic system that works based on the personal device handshaking and body tracking [2]. Wang et al. introduced Proxemic Peddler, which is a public display that can capture and preserve the attention of pedestrians [69]. All of the aforementioned studies analyzed proxemic interactions at the micro-level where the proximity space is very small (e.g. one room).

Exploring the macro-level of proxemic interaction has been a popular research subject for the past few years. As such, researchers have been working on designing proxemic interaction analyzing systems for larger spaces (e.g. multiple rooms). One of the first and most famous examples is the Active Badge tool [70]. In its related study, Want et al. introduced a tracking system via a wearable badge. An infrared beacon was embedded into the badge and connected to a sensory network. The sensory network updated the location of the badge every 15 seconds by picking up the signal emitted by the badge. Want et al. originally designed the Active Badge to be used on a large scale. After they successfully implemented it in an office to redirect employees' phone calls to their latest known stations, Active Badge was deployed in several other locations.

Laibowitz et al. proposed another proximity-aware system in the work space called UbEr-Badge [34]. This system uses another type of wearable badge that can detect stationary location tags or other badges. The main purpose of this system is to facilitate social interaction in large and crowded locations (e.g. large meetings, workshops, conferences, etc.). Some of the applications of this system include finding other people through their badges and wirelessly exchanging contact information.

There exist some studies that have been focused on proxemic interaction in the home area. The EasyLiving project by Brumitt et al. is one of the earliest studies in this category [11]. In their research, they placed the emphasis on architecture of the system, which can collect and connect different devices together in order to enhance the user's experience. They made this possible by tracking all devices and users using various technologies and binding them using a middleware application. In the same vein of thought, Ramani et al. proposed a new location tracking system for media appliances that makes use of wireless home networks [50]. Though their study's main goal was to implement a tracking system while they mentioned few practical applications of the system they created.

Among various applications, we can name the proxemic media redirection system, which

Table 2.1: The comparison between ProGes and reviewed studies.

Method	Content Type	Device Selection Mechanism	Control User Interface
Active Badge [70]	Audio	Tracks the employee and selects the closest station	—
EasyLiving [11]	Audio/Video	Manual	Graphical
Ramani [50]	Audio/Video	Automatically selects the closest device.	Graphical
AirPlayer [61]	Audio	Automatically selects the closest device.	Graphical
ProGes	Audio/Video	Recommends a device based on distance, user profile, and device and content properties. User can accept or reject the suggestion and it learns user's attitudes over the time.	Graphical and Natural (Gesture)

is supposed to redirect the media stream from one device to another using proximity sensing. Recently, Sørensen et al. introduced the AirPlayer which is a multi-room music system that uses proxemic interaction [61]. AirPlayer is a self-positioning system that allows each device to find its location and act independently [17]. As a result, each AirPlayer device first finds its location by measuring and comparing the Received Signal Strength Indicator (RSSI) for each pair of devices. Then, if there is an Apple's AirPlay device available nearby, a request for media stream redirection is sent to from the AirPlay device. Furthermore, they conducted a field evaluation with promising results on user satisfaction.

In conclusion, we have reviewed a few proxemic interaction studies in this section. Table 2.1 presents a comparison between this study and reviewed studies to highlight the differences and advantages of ProGes (our proposed UI). The first column shows the supported content types. As we can see, AirPlayer only supports audio while the rest support both audio and video contents. The second presents how a device is selected by each system. In EasyLiving, the user selects a device manually. AirPlayer and Ramani's method, however, only need the distance information to be able to redirect a media stream. That being said, ProGes has a cloud-based decision maker that considers users' profiles, content properties, and devices' capabilities in addition to distance information in order to recommend a device to the user. Then, using the feedback from the user, it may or may not start a media redirection interaction. Furthermore, ProGes learns the habits of the user and updates itself. The third and last column compares the type of user interface for these 4 systems. While the first 3 systems only provide graphical user interfaces, ProGes has the extra feature of NUI allowing it to be implemented by hand gestures.

2.2 Gesture Interaction

Gesture interaction is a very broad area of study that can include different research fields, several technologies and various techniques. For example, gesture interaction is not pos-

sible if we cannot automatically recognize the gestures. As a result, a review on gesture interaction should include both gesture recognition and gesture elicitation studies. This section first presents a brief background on gesture recognition. This also includes the implemented method for an evaluation study of the proposed vocabulary. Then, a literature review on gesture elicitation studies is made to show the differences between our proposed methodology and related work.

2.2.1 Background

Due to breakthroughs in touchscreen and computer-vision technologies, research in gesture interaction has been highly active in the recent years. Gesture interaction systems now support a wide variety of input types (e.g. RGB cameras, IR cameras, etc.) and different number of sensors (e.g. monocular vision, stereovision, multi-camera, etc.). However, the most important aspect of these systems is that they can extract depth information from the scene in real-time. Using depth information, we can recognize and track hands, body parts, and gestures more accurately. We will shortly review vision-based hand/body gesture recognition and tracking in this subsection.

Vision-based hand and body recognition include similar steps. The first step is extracting shape features from the input image. There already exist various hand recognition algorithms that depend on the geometric features of the hand such as the finger and palm heights, finger shapes, palm prints, hand contours, etc. [77]. Following the collection of these data, a template-matching algorithm classifies them in order to create signature representations. Finger-Earth Mover's Distance is an example of such an algorithm and uses a distance metric for hand dissimilarity measure [52]. Breuer et al. proposed a similar approach however they used 3D point clouds of the scene instead of a template matching to fit an articulated hand model in order to find the position and orientation of the hand [8].

Bai and Latecki introduced the skeleton graph matching, which uses object silhouettes [5]. Their method is considered as a successful approach for human body recognition. Zhu et al., on their end, proposed a flexible Bayesian framework for human pose estimation [80]. Their method is based on the optimization of anatomical landmarks of the human body. More recently, Shotton et al. introduced real-time human pose recognition, which works with no temporal information (i.e. it only uses a single depth image) [60]. This method extracts human body parts and computes 3D joint positions to recover the whole body configuration.

Researchers have been constantly developing tracking algorithms in various contexts since 1980s. Hoffman et al. proposed solution for computing the 3D motion of animal limbs from the 2D motion of their projected images. Their solution used the anatomical constraints in the analysis of biological motion [25]. Lee and Kunii studied a model-based approach to detect hand posture by using its biological constraints [36]. Later on, a new motion modeling method that combines local and global representations of human body parts to quantize them in polar space was proposed by Tran et al. [64]. Another common approach for human body analysis is tracking body parts with Kalman filters [54, 75].

However, it must be noted that the latest (and most popular) approaches tend to use particle filters instead [16, 28].

Gesture recognition techniques (i.e. the combination of recognition and tracking) can be used in different application domains. Therefore, they can be very simple or quite complex depending on the application requirements. For example, they can be as simple as tracking the movement of fingers to draw a line (e.g. FingerPaint [15]), or they can be more sophisticated and use advanced gesture recognition techniques like sign language interpretation [6, 63, 68, 76].

The most advanced gesture recognition techniques are the ones used in Robotics, Virtual Environment, and Device/Application control where the gestures are more complex. Hidden Markov Models (HMM) is one of the famous techniques for gesture recognition in these domains. Studies such as [14, 48] used HMM to recognize dynamic gestures (e.g. waving your arm) from image stream and transform them into direct robot motor commands for arm or whole body imitation. Similarly, Lee and Kim presented an HMM-based recognition system to control a presentation in a slide show presentation application [35]. The other famous gesture recognition technique in these domains is Dynamic Time Warping (DTW) [53]. Celebi et al. proposed a weighted DTW algorithm that gives joints different weights based on the importance of joints in each gesture [12].

In conclusion, this subsection presented a background and review of vision-based hand gesture recognition, which is a broad and highly active research field. However, this area of recognition does not provide the main contribution to this study. In the following pages, we will propose a gesture vocabulary for multimedia devices based on the user's preferences. In order to validate the proposed vocabulary, we will explain how we adapted a developed DTW algorithm to conduct a user evaluation study.

2.2.2 Hand Gesture Elicitation

Users are always considered as the primary priority in the human-computer interaction point of view. Due to this priority, researchers have developed a new method for product and system design, which is now very common. In this method, users are asked to answer some questions and give feedback. Then, using the collected data, researchers emit different rules and guidelines for the design process. Some researchers believe this method is the basic idea behind the participatory design [57].

In the last few years, this methodology has reached into the domain of interactive technologies. Gesture interaction is very popular in several technological fields making use of a non-traditional interaction medium. HCI systems like mobile, surface or tablet and hands-free computing have developed and embedded gestures that are proposed by expert system designers. Yet, as we can see in the results of numerous studies, a gesture vocabulary that is designed by HCI experts may have been manipulated by the characteristics and limitations of gesture recognition engines [74].

As a result, in 2003, Nielsen et al. delineated a procedure for designing gesture vocabularies based on hands-free computer interaction in the ubiquitous computing that considered the users' point of view regarding intuition, learning rate and ergonomics [47]. Two

years later, Wobbrock et al. accompanied this research with a preliminary user-elicited study on gesture vocabulary design, which targeted unistroke gestures [73]. This was the first time that the concepts of guessability and agreement scores were introduced. Almost all of the subsequent user-elicited studies also made use of these concepts.

We conducted a survey on the latest user-elicited studies that are relevant to the goals and scope of our research. All of them aimed for optimal gesture vocabulary design and were compared based on their main characteristics (i.e. methodology, interaction media, number of commands, etc.). Table 2.2 presents the conclusion of the comparison between our research and the works reviewed above.

The present survey starts with a review on research by Webbrock et al., which proposed a user-elicited gesture vocabulary for surface computing applications using the Microsoft Surface prototype [74]. In their study, they demonstrated the results of a comparison between two gesture vocabularies. The first one was proposed by three HCI experts while the second one was elicited from a user study. Their results indicated that only 60% of user-elicited gestures were the same as experts' vocabulary while 19.1% of experts' suggested gestures were never tried by participants. Moreover, they continued this study and reached the conclusion that user preferences tend toward simpler physical and conceptual patterns for the gestures [44]. Thus, these findings make it clear that involving the users in the development of interactive technologies can offer advantages that may lead to greater level of future appreciation from the consumers.

Nacenta et al. conducted a study based on another aspect: Memorability [45]. They presented a comparative study between user-defined versus pre-defined gesture sets by evaluating their memorability to corroborate these findings. Indeed, the key verdict in that study was that individual user-defined gestures are easier to remember than pre-designed gestures for individuals. This highlights another highly beneficial aspect of user-elicited studies.

Following this trend, Ruiz et al. conducted a study on a user-defined motion gesture set for mobile devices to elicit a vocabulary of device movements that invokes a set of general commands like answer phone, ignore call, navigate mail, etc. [55]. Although they studied a different medium and context, there are some similarities between their vocabulary and ours. This indicates that there is a strong intuition connection between concepts and some gestures that can dominate the context and boundaries of the technology.

There exists a great number of gesture elicitation research concerning mobile devices but we want to highlight a particular study because of its specific target. Seyed et al. aimed to propose a gesture set for data transfer tasks in multi-display environments (e.g. a touchscreen tablet, a touchscreen tabletop, and a wall display together) yet they failed to derive a final gesture set [59]. Several complications stopped them from reaching to their original goal, the main one being the users' lack of familiarity with multi-display environments.

Of successful domain elicitation studies, we have reviewed research relating to both surface and mobile computing applications as well as hands-free interaction. Since the release of the depth-based sensors to the mass market, in particular the Microsoft Kinect sensor, hands-free interaction by means of gesture recognition has become very popular.

Table 2.2: The comparison of user-elicited studies for gesture set design.

Study	Interactive Media	Participants	Demographics	Commands (referents)	Methodology
Wobbrock et al. [74]	Surface computing *(Microsoft Surface prototype)*	20	Non-expert users, never used a touch-screen device before.	27	Demonstrative (Participants see the simulated effect of a referent and ask to perform a corresponding gesture). *Format: Interview*
Ruiz et al. [55]	Mobile computing *(Multi-platform)*	20	Non-technical high-tech company workers.	19	Inquiry (Participants designed and performed a motion gesture for a specific referent). *Format: Interview*
Seyed et al. [59]	Multi-display environments (Apple iPad, Microsoft Surface 2 and SMART Board Display)	17	Variety of backgrounds (no exclusion based on experience)	16/device	Same as [74]
Vatavu [65]	Free-hand gesture recognition *(Microsoft Kinect sensor)*	12	CS students with no experience in interaction design or using a Kinect.	12	Demonstrative (Participants read and watched a video of the effect of a referent and ask to perform a corresponding gesture). *Format: Interview*
Morris [43]	Free-hand gesture recognition and speech recognition *(Microsoft Kinect sensor)*	25	Varied professions, wide age-range, prior experience using Kinect	15	Same as [74]
Connell et al.[13]	Free-hand gesture recognition *(Microsoft Kinect sensor)*	6	Children age 3 - 8 with experience using touch-screens and smartphones.	22	Same as [74]
Piumsomboon et al. [49]	Free-hand gesture recognition *(Sony head mounted display, HD webcam and Asus Xtion depth sensor)*	20	Participants with minimal knowledge of AR. Experience with PCs, touch-screens, Wii and Kinect	40	Same as [74]
Vatavu [67]	Hand-held motion sensing *(Wii Remote Controller)*	20	From technical to research level backgrounds.	12	Same as [65].
Proposed in this thesis	Free-hand gesture recognition *(Microsoft Kinect sensor)*	49 in preliminary study, 81 in online questionnaire, 20 in technical evaluation	People from different background in the preliminary study. University faculty and students from various departments for the online questionnaire and technical evaluation.	12	Two-phase Online Inquiry (Phase 1: Participants suggested a mid-air hand gesture for a specific referent, Phase 2: Participants selected the most preferred gesture from suggestions of phase 1). *Format: Online-survey* User agreement and memorability test to validate our results. *Format: Lab-based user study*

We begin our review of studies in this domain with Morri's study [43]. She explored users' preferences for web browsing using a big screen TV in the living room. In order to conduct this study, she used the Microsoft Kinect sensor to combine gesture and speech recognition together and build a new navigation mechanism for TV web browsing. Her findings can be categorized into two items: first, participants liked the idea of web browsing on TV for some scenarios and applications. Second, multiple modalities (i.e. gesture and speech) increased the guessability of the commands for the participants.

Vatavu also conducted a series of studies in this domain. He began his preliminary work with the intention of designing a gesture set for TV control tasks [65]. Then, he continued his work by adopting this methodology in the area of Augmented Reality (AR) [67]. He took Nintendo's Wii remote control to develop a set of interaction commands for a prototype hybrid-physical augmented TV whose multimedia contents span the physical space of the TV. Last, he extended his work by putting those two studies together and composing a comprehensive comparison of hands-free gestures versus handheld gestures [66]. He concluded his work by proposing two gesture vocabularies - one for each type of interaction - and discussed the advantages and disadvantages of handheld and hands-free interactions.

Likewise, there are few more studies in this field that do not have the same goal as ours yet be mentioning as they modified the pre-existing methodology mentioned above. Piumsomboon et al. suggested a user-defined gesture set for 40 AR tasks such as object manipulation, object transformation, editing, menu navigation, etc. [49]. A more recent study focused on child-defined gestures [13]. Connell et al. believed that child-defined gestures are an equally important yet ignored area because children's gesture performance and preference is not similar to adults as they, unlike their adult counterparts, tried to use whole-body gestures for object manipulation, navigation-based tasks and spatial interaction. It must be noted that despite their best efforts, Connell et al. could not conclude their work because they were unable to obtain an average level of agreement for any given command.

Table 2.2 presents an abstract comparison of reviewed user-elicited studies. Most of these studies adopted the same methodology originally introduced by [73] and which can be described as demonstrative. As will be illustrated in the following chapters, we are proposing a new methodology, which yields to good agreement levels. In the inquiry-style online survey we used to test our methodology, we gathered information concerning not only suggested gestures but also regarding preferences and attitudes towards interactive hand gesture technology used to control multimedia devices and applications. In addition, my methodology has three parts that are not present in previous studies: A technical evaluation, an external user agreement experiment and a memorability test.

Chapter 3

User Study

The proposed system ProGes is a multifaceted system that is designed based on user attitudes and preferences. As such, before we design and introduce our proposed system, we need to first study users concerns and desires. Only then we are able to build the ProGes, which is described in the next chapter. This chapter aims to present the description and results of a series of user surveys. We conducted these surveys in order to obtain information on the thoughts, feelings, and favorites of users for gesture controls and multimedia devices. There are two sections in this chapter: Gesture Elicitation and Multimedia Device Preference Elicitation. In the first section, Gesture Elicitation, we give a detailed explanation on a two-step user survey that helped us design our gesture control unit. The second section explains the details of our second survey, which was conducted to find the user preferences on multimedia devices.

3.1 Gesture Elicitation Study

People are eager to learn and experience the cutting edge technologies and interaction methods that have recently begun becoming available for public consumption. A good example of this is shown by the popularity of touch-screen devices such as smart phones and tablets. As chapter 2 showed in a review on the active research in gesture recognition and elicitation studies, Microsoft Kinect and similar products have greatly influenced and improved the field of gesture interaction. However, since gesture interaction is an advanced and complex research area, there is always a demand for a higher user experience quality. This demand has motivated us to focus on building a vocabulary that is inferred by the user's preferences, as seen in the first part of this study. Yet, designing a gesture vocabulary is a complex activity.

Questionnaires are one of the favorite methods for studying user preferences. As such, we decided to develop a questionnaire to target user preferences in terms of an interactive hand gesture vocabulary. In order to design that questionnaire, a set of control commands should first be chosen (as well as the effect of those gestures). These control commands are called *referents* in linguistics, and because Vatavu [65] and Wobbrock et al. [74] used

the same terminology, we follow the popular motion and also call them referents in this study. Since multimedia devices and objects are the main target of the ProGes, a set of referents that can be applied to televisions, digital displays, gaming consoles, etc. is selected. This set has 12 different control commands, all of which are frequently used by multimedia devices. These referents are categorized into four classes: Menu Navigation, Channel Surfing, Volume Control, and Override. Table 3.1 presents these commands.

Table 3.1: Referents that are used for the design of our user-elicited gesture set.

Group	Referents	Description
Menu Navigation	Open	Open menu
	Up	Move up in menu
	Down	Move down in menu
	Enter Option	Enter option in menu
	Enter Submenu	Enter a submenu
Channel Surfing	Previous	Go to previous channel
	Next	Go to next channel
	Last	Go to last visited channel
Volume Control	Increase	Increase volume
	Decrease	Decrease volume
	Mute	Mute volume
Override	Shutdown TV	Shutdown TV

Furthermore, there is an import point that should not be ignored when users are involved in the development process. While users may not limit themselves to the boundaries of present technology, they can be biased due to their experience with available devices. In order to overcome this issue, the proposed methodology uses a two-phase user survey wherein one informs the other. The results of this survey are used in the development of the ProGes and assessed by a technical evaluation and external user agreement experiment revealed in the following chapters. We believe that this methodology can help us achieve higher agreement scores for the proposed gesture vocabulary.

3.1.1 User Survey

The two phases of the proposed methodology are called Preliminary Study and Online Questionnaire. Although these two phases share the same questions, they have different goals and answers. The first phase is an open questionnaire to collect as much information as possible while the second phase is a multiple-choice questionnaire to narrow down the answers and compose a vocabulary. This subsection explains the details of the participants and procedure.

Participants

The preliminary study phase targeted a general audience such as friends, acquaintances, colleagues of the researchers, etc. in order to consider a wide range of people with different backgrounds. 49 people participated in this phase of the user survey. It should be

mentioned that the average age of participants was between 30 and 40 years and they originated from different regions around the world like North America, the Iberian Peninsula, the Middle East, etc.

In the second phase of this user survey, there were 81 participants with an approximately equal gender distribution (44 male and 37 female). The second phase was run among university students and staff. In this phase, participants were asked to indicate their age range from a list of predefined age intervals. The average age was between 22 and 26 years (sd. = 6.7), which was not a surprise because of the nature of participants' domain. Table 3.2 presents the correlation between the age and gender of the people who participated in the second phase of the survey. Moreover, since we wanted to separate the learning and testing groups, participants of the first phase did not take part in the second phase. As a result, these two groups were completely independent.

Table 3.2: The age/gender distribution of participants in the second phase of user survey.

Gender / Age Distribution (%)									
Female					Male				
18 - 21	22 - 25	26 - 30	31 - 40	>40	18 - 21	22 - 25	26 - 30	31 - 40	>40
23	9	4	7	2	15	14	7	12	7

In addition to their age and gender, the second phase asked participants about their video game playing frequency. Our predictions concerning the answer to this question was that there would be more users with higher playing frequency in the lower age range however the results did not confirm this prediction. Participants from the youngest age group (i.e. 18 - 21) mostly answered that they play "sometimes, rarely" (12.35%) or never (24.69%). Furthermore, across all age groups, people who never play (27.16%) or rarely pay (34.57%) composed the majority of participants. This can be interpreted as a good sign since frequent video game players could be biased towards certain gestures for some types control commands. Moreover, participants of this phase came from different departments of the university. As a result, this phase covered a wide spectrum of technical backgrounds, which prevented any unwanted influence of background knowledge.

In summary, since the majority of participants were not frequent video game players and they had different technical backgrounds, we find that the answers acquired from this user study can accurately represent our targeted users of a more generalized demographic.

Procedure

Participants were asked to answer the same questions in both phases of this user survey although their answers had different characteristics (i.e. open versus multiple-choice questions). The questions were developed in such a way to be as general as possible in order to elicit user preferences towards a hand gesture vocabulary that is suitable for wide range of applications and users regardless of their age, gender or technical background. Both phases of this study were conducted as online user surveys wherein participants had to complete the questionnaire's form and submit it.

Participants of the first phase were asked open-ended questions. The main goal of this phase was to compose a list of possible gestures for each referent as well as data collection. There were two important points in this phase of study. First, participants were asked to not assume the current state of gesture recognition technologies. Second, they were not provided with any examples or feedback during the study. These assumptions were considered in order to address a very well-known problem in elicitation studies. This solution is called removing the Gulf of Execution and it is mentioned by similar studies such as [55, 74]. Furthermore, this open-ended questionnaire gave participants an opportunity to use their imaginations without any limits. Participants were also encouraged to give more than one answer to each question whenever possible.

The second phase of this study was a multiple-choice questionnaire. Answers of the first phase were used to develop the second questionnaire. If there was less than or equal to five different answers for a question in the first phase, all of them were taken into account as the options of the question in the second phase. For questions with more than five different answers, the top four or five most frequent answers were chosen as the options of the second questionnaire. The primary intention behind this two-step procedure was that the final results, no matter what they were, would be completely user-oriented. Again, neither examples nor feedback were provided to participants during the second phase in order to prevent any potential impact on their answers from the researchers.

3.1.2 Results

Although the participants in the different steps of this study were completely separated from one another, the presented methodology follows a single thread. Hence, the results achieved by each step formed the base of the phase to follow it. This section presents the results of each step and gives an explanation on how these results affected the ones to follow them.

Preliminary Study

The goal of the preliminary study was to collect data and create a vast set of gestures for the studied commands. To achieve that goal, the participants were asked to describe as many gestures they think of for each command. Table 3.3 shows the results of this preliminary study.

In this phase, 49 people participated in the study and 34 sets of answers were valid since the others were either incomplete or vague. As is shown in Table 3.3, the total number of suggestions for each command is always greater than the total number of participants (34 people). Since participants of this phase of the user survey were not limited to the boundaries of current technologies, they could freely use their imagination to suggest different gestures. As a result, they proposed gestures with various degrees of freedom in 2D and 3D space that used different parts of the body such as hands, ears, mouths, etc. In total, 511 gestures were suggested by participants through this preliminary study (an average 42.58 suggestions per command). We grouped similar gestures together and counted them

Table 3.3: Proposed gestures for selected control commands of multimedia devices.

Referent	Total No. of Suggestions	No. of Different Gestures
Next Channel	44	11
Previous Channel	45	11
Last Visited Channel	42	16
Turn Up Vol.	44	12
Turn Down Vol.	46	13
Open Menu	45	15
Open Submenu	40	18
Up	42	12
Down	41	12
Enter Option	41	19
Mute	42	16
Shutdown TV	39	22

as one. A good example of this grouping procedure is for the "Shutdown TV" command where some participants suggested clapping once while some others mentioned clapping twice. Thus, they were grouped together and we created the clapping option for this command. The third column of Table 3.3 shows the number of different groups of gestures for each command. "Shutdown TV" had the most variety of suggestions with 22 gestures and "Next/Previous Channel" had only 11 distinctive suggested gestures. On average, 14.75 distinct gestures were anticipated by participants for each command.

Since our participants suggested at least 11 different groups of gestures for each command (Table 3.3), we had the opportunity to choose the possible answer choices for the next phase of the user survey. We decided to limit the number of choices to 4 or 5, similar to the number of choices in [44]. Later, this restriction helped us to achieve a good agreement score level on the second phase where we had more than 80 participants. Gestures mentioned with the highest frequency in the first phase were selected as the answer choices of the second phase of the user survey. Furthermore, it should be mentioned that we did not limit the choices to one-handed or two-handed gestures. Thus, we had choices like clapping in the second phase.

Online Questionnaire

This subsection presents the results of the second phase of the user survey. It also shows the elicited gesture vocabulary, which is then analyzed using the agreement score test. Moreover, a more detailed statistical analysis was conducted to extract the correlations and dependencies between the answers. Results and findings of the statistical analysis are given in Appendix A. Table 3.4 presents the second phase's questionnaire and the distribution of results.

Table 3.4: The questionnaire of interactive hand-gesture vocabulary. Percentages in bold represent the preferred answers. Underlined percentages refer to the highest preference.

Index	Question and Answer Options				
Q1	Select the age range that you are in:				
	< 18 - 21	22 - 25	26 - 30	31 - 40	> 40
	38.27%	22.22%	11.11%	19.75%	8.64%
Q2	Please select your gender:				
	Male		Female		
	54.32%		45.68%		
Q3	Do you usually play videogames?				
	Almost every day	More than once a week	More than once a month	Sometimes, rarely	Never
	8.64%	13.58%	16.05%	34.57%	27.16%
Q4	How often do you play videogames in which the game control is made using body movements?				
	Almost every day	More than once a week	More than once a month	Sometimes, rarely	Never
	4.94%	1.23%	2.47%	43.21%	48.15%
Q5	Your gestures: Next Channel				
	Hand/Arm from left to right / right to left	Head to the left / right	Eyes to the left / right	Hand to the front	
	88.89%	2.47%	3.70%	4.94%	
Q6	Your gestures: Previous Channel				
	Hand/Arm from left to right / right to left	Head to the left / right	Eyes to the left / right	Hand to the back	
	83.95%	2.47%	3.70%	9.88%	
Q7	Your gestures: Back to the last visited channel				
	Hand/Arm from left to right / right to left	Hand / Arm from top to bottom / from bottom to top	Hand to the back	Circular movements with the hand	
	1.05%	14.81%	23.46%	45.68%	
Q8	Your gestures: More volume				
	Hand/Arm from left to right / right to left	Hand / Arm from bottom to top	Thumbs up	Circular movements with the hand	
	6.17%	49.38%	28.40%	16.05%	
Q9	Your gestures: Less volume				

	Hand/Arm from left to right / right to left	Hand / Arm from top to bottom	Thumbs down	Circular movements with the hand	
	3.70%	49.38%	30.86%	16.05%	
Q10	Your gestures: Open menu				
	Hand to the front	Make a fist with the hand	Clap / Unclap hands	Circular movements with the hand	Snap Fingers
	28.40%	14.81%	17.28%	11.11%	28.40%
Q11	Your gestures: Menu, enter a submenu				
	Hand/Arm from left to right / right to left	Open/Close the fist	Clap / Unclap hands	Point at the TV	Stop sign
	13.58%	34.57%	8.64%	39.51%	3.70%
Q12	Your gestures: Menu, up in the menu				
	Hand/Arm from left to right / right to left	Hand / Arm from bottom to top	Thumbs up	Point at the TV	
	9.88%	48.15%	28.40%	13.58%	
Q13	Your gestures: Menu, down in the menu				
	Hand/Arm from left to right / right to left	Hand / Arm from top to bottom	Thumbs down	Point at the TV	
	8.64%	48.15%	27.16%	16.05%	
Q14	Your gestures: Menu, enter an option				
	Push and point	Fist and point	Yes with the head	Thumbs up	Snap Fingers
	56.79%	17.28%	16.05%	3.70%	6.17%
Q15	Your gestures: Mute / Unmute				
	Hand/Arm from left to right / right to left	Hands on the ears	Hand on the mouth	Open/Close hand (without bending the fingers)	Make an X with the arms
	4.94%	20.99%	28.40%	30.86%	14.81%
Q16	Your gestures: Shut down the TV				
	Clap hands	Hands on the eyes	Stop sign	Bye gesture	Make an X with the arms
	19.75%	8.64%	17.28%	39.51%	14.81%
Q17	Do you prefer accuracy rather than speed (time response)?				
	High	Moderate	Intermediate	Almost nothing	I don't care
	24.69%	33.33%	34.57%	4.94%	2.47%

Q18	Do you like the idea of a set of recognition gestures instead of conventional commands?				
	High	Moderate	Intermediate	Almost nothing	I don't care
	17.28%	33.33%	25.93%	16.05%	7.41%
Q19	Would you remember a set of gestures for easy navigation on TV?				
	All of them	Most of them	Half of them	Some of them	None of them
	23.46%	48.15%	9.88%	14.81%	3.70%
Q20	Would you find it more useful to have one gesture for each option (enter menu, enter volume options, enter channel options...) and repeat the ones for up/down?				
	Yes	No	I don't mind		
	74.07%	13.58%	12.35%		
Q21	Would you prefer a system controlled by gestures rather han a conventional remote control?				
	Yes	Only a set of choices	Both	No	
	29.63%	20.99%	33.33%	16.05%	
Q22	Why do you choose a particular gesture to an associated control command (e.g. Volume Up)				
	It's more intuitive	It reminds me to the smartphone	It's more comfortable	It is based on standards	For no special reason
	67.90%	6.17%	13.58%	6.17%	6.1%

The proposed interactive hand gesture vocabulary is extracted using the gestures with the highest frequencies in the results. In some cases, two gestures were too similar so we selected both of them in this step. Those cases are discussed in greater depth in the evaluation and discussion to follow. Figure 3.1 presents the proposed hand gesture vocabulary for controlling multimedia devices. Correlations between different referents can be seen in Figure 3.1, for example Increase/Decrease Volume and Up/Down in the menu. Furthermore, some similarities to multi-touchscreen gestures are identified in referents with linear horizontal/vertical gestures as well as pointing and pushing. However, for referents that have a specific button on a touchscreen device such as Mute/Unmute or Shutdown, the interactive hand-gestures suggested by the users are quite interesting, since they resemble body/sign language (i.e. placing a hand over ones mouth for Mute/Unmute or a wave goodbye for Shutdown).

The extracted gestures for each referent were evaluated by computing an agreement score A_r, which represents the level of consensus between the participants for a specific referent r. This agreement score was initially introduced by [73] and it is used to obtain agreement scores of similar user-elicited studies. The agreement score A_r for a specific referent r is calculated as follows:

$$A_r = \sum_{P_i \subset P_r} \left(\frac{|P_i|}{|P_r|} \right)^2 \tag{3.1}$$

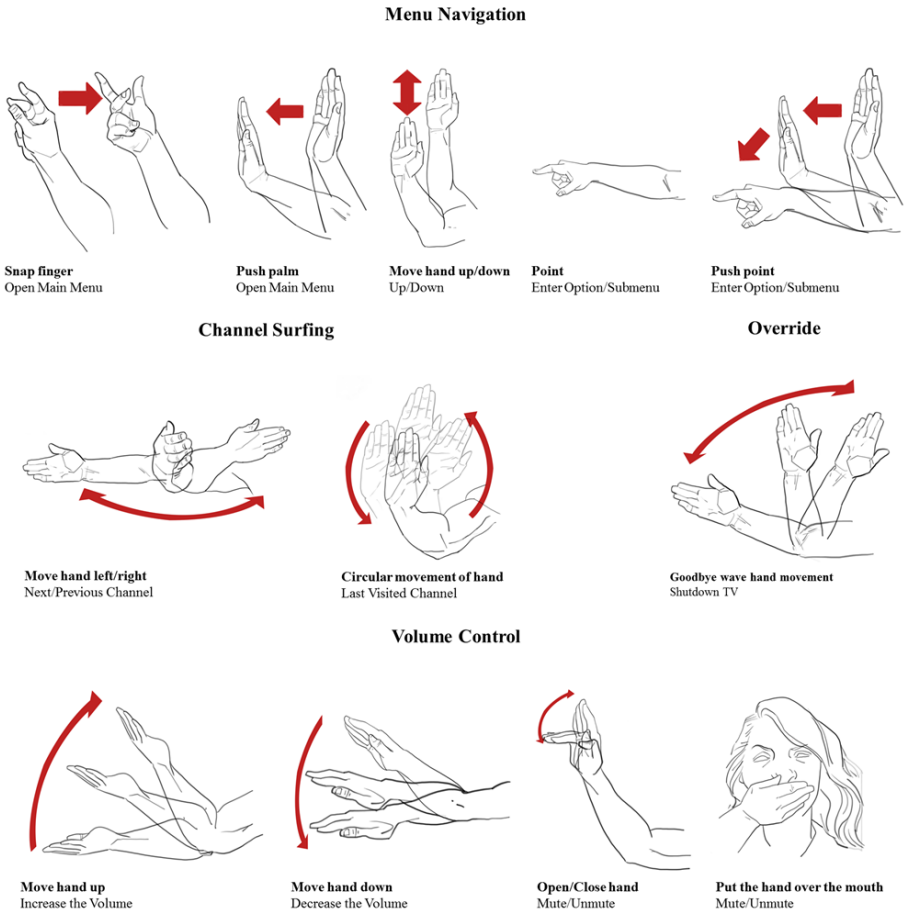


Figure 3.1: The elicited hand-gesture vocabulary.

where P_r is the set of proposed gestures for the referent r and P_i is the subset of identical gestures within P_r. The value of A_r ranges between $|P_r|^{-1} \leq A_r \leq 1$, where $|P_r|^{-1}$ represents no agreement at all and 1 is perfect agreement. The studied number of proposals is the result of 81 (number of participants in the second phase) $\times$ 12 (referents) = 972 proposals. In this evaluation, P_r has the same value as the number of users in the second phase because that is the studied training set value and dismisses incomplete surveys.

The main references of related gesture elicitation studies are [65, 67, 74] and so we compare the results of this study with them. In the beginning of the results analysis, an indication of a higher divergence can be seen in our responses when compared to the aforementioned studies. This is due to two reasons: First, the number of participants in this study 81 people is almost 4 times larger than that of similar studies (20 people in [65, 67, 74]). Second, neither examples nor feedback were provided to participants while similar studies provided at least one of them. The hired elicitation method is completely based on the user preferences and imagination in both the proposition and selection of referents. There is no external influence whatsoever in the whole process.

Although the proposed method in this study is different from [65, 67, 74], their results can be referenced by the present study to see how effective the proposed method is. Since the present study has more participants, it was expected to have more distributed results. However, the results obtained tend towards the opposite of the expected ones. The mean agreement score for the proposed set is 0.56, which is higher than that of similar studies. The agreement values obtained by [67] is 0.53 for hand-held motion control within an augmented TV prototype. It is 0.42 for free-hand TV control in [65]. Wobbrock et al. achieved 0.32 for surface computing [74] and the agreement value equals to 0.28 for mobile phone interaction in [55]. The agreement rates were observed for each suggested referent and the average agreement rate of the proposed vocabulary are presented in Figure 3.2. The standard deviation is 0.12 (0.47 in Wobbrock et al. [74] and 0.14 in Vatavu [67]). Focusing on the graph, it can be observed that the gestures with highest agreement rates are the ones related to pointing and moving the hand sideways or up/down.

3.2 Multimedia Device Preferences Study

Advances in technologies have dramatically changed our lifestyle in recent years. We interact with various kinds of devices everyday as we are surrounded by a large number of them in the IoT. For example, just a decade ago, there were not too many choices when deciding which tool to use to play a song or video. However, a diverse group of media player devices is now available in today's market. Although multimedia devices are only one group of devices in the IoT, this group contains a large number of items that have a huge impact on our personal life. Multimedia devices use either different user interaction technologies (e.g. touchscreen, handheld controllers, etc.) or dissimilar commands (e.g. different gestures for the same referent). As such, we can claim that they are completely

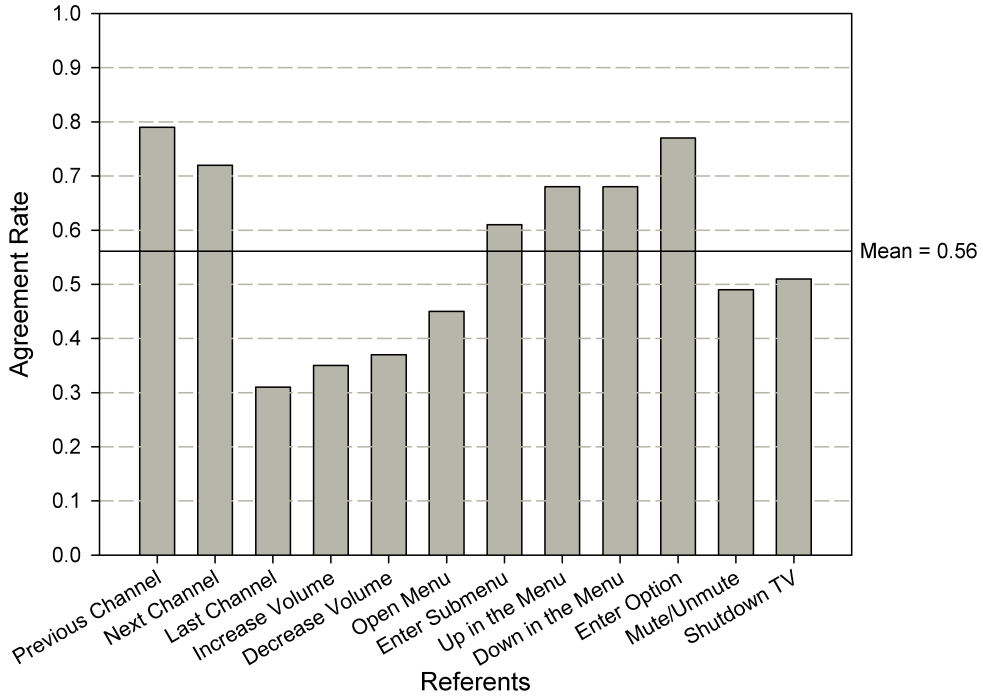


Figure 3.2: The agreement score test values for user-elicited hand-gesture commands.

independent nodes in the IoT. Yet, we believe that we can improve users' experiences by proposing a new UI that connects devices together and provides a single set of control commands for them. In order to do that, we decided to conduct a survey on users' attitudes and preferences concerning multimedia devices. In this study, we asked questions about how much time a user spends on watching or listening to multimedia contents everyday, users' priorities and preferences on multimedia devices and their expectations from a dream multimedia system.

3.2.1 User Survey

We developed a non-technical questionnaire for this survey due to two important reasons. First, people usually choose a multimedia device based on the media content and the situation in which they find themselves. This creates an extremely large combination of scenarios; we could not possibly have hoped to cover all of them. Therefore, we had to ask general questions. Second, asking questions that need a minimum level of knowledge or background limits the participants to a group of people who have some common properties. This restriction biases the survey results, which is in contradiction with the desire to elicit users' general preferences. Thus, questions were designed to be simple and clear for everyone. In addition to that, we added well-known examples to some questions where an example could make the question clearer (e.g. we provided iPhone as an example for

smartphones).

Participants

This user survey was conducted in one step. Also, we initially targeted to have at least 100 participants answer the survey. At the end of this survey, a total of 149 participants had answered the questionnaire. Moreover, there were three independent demographic factors that we had to consider when selecting participants in order to ensure that our survey has unbiased results: age, gender, and job type.

Regarding the age, participants were asked to specify their age range among five options: 10-19, 20-29, 30-39, 40-49 and more than 50. If we assume the median of each age range as being the representative, the total average of age is 28.62 years old. People of this approximate age commonly try new technologies and devices. As such, it was a benefit for us as most of the participants were familiar with multiple kinds of devices and technologies and, as a result, they could answer our questions. However, there is a drawback. Our general results could be biased by the opinions of this age demographic. To mitigate this possibility, we studied results based on the age range. We will explain the age-based analysis in greater detail in the results section.

Our second concern was about the gender of participants but we can claim that the overall results of our survey were not overly influenced by gender differences. Indeed, the numbers of male and female participants were almost equal (53.7% male participants and 46.3% female participants).

Our final concern was related to the job type of the participants. We asked them whether their job is related to IT or not. Since there is no unique definition for IT job, there is always a debate on what does or does not fit into this category. Due to this, we did not provide a definition for "IT-related job" and trusted the participants' judgment. All in all, our participants were divided into almost equal groups. 53% of them had IT related jobs and the rest worked in non-IT related jobs. Consequently, we can claim that the overall results are not biased based on the participants' job type.

Procedure

The questionnaire was designed to be conducted anonymously in order to prevent any privacy violation issues. It had 17 questions that can be categorized into one of three groups:

1. **Demographics:** The first three questions belong to this group; we asked participants about their age, gender and job type.
2. **Attitudes:** The second group also has three questions, these being about how much time each participant spends using various media contents at home and while they are driving.

3. **Preferences:** The last group of questions was the most important part of this survey for us because it presented the questions related to user preferences on media playback devices and future systems.

When we carry out a survey, we want to see unbiased and real results. Thus, questionnaires should not have influence participants in favor of its designers' intentions. In order to keep our survey unbiased, participants were not told about the details of the research. Furthermore, we did not provide any example answers for participants. In this survey, the main targeted audiences were ordinary people so our questionnaire had to be accessible to everyone in the same way. Hence, we decided to conduct it as an online survey. We used Google forms to prepare the questionnaire. The majority of the questions were in multiple-choice format and the rest were drop downs. Then, we uploaded it onto the cloud: Google drive in this particular case. Our next step consisted of publicizing the link and sending it to our friends and acquaintances, which allowed a broad range of people to have access to the form.

3.2.2 Results

Table 3.5 presents all of the questions and the distribution of answers for each question. Answers are discussed and analyzed in the following two subsections: Answers, and Statistical Analysis.

Table 3.5: The questionnaire and the distribution of answers.

Index	Question and Answer Options				
Q1	Please specify your age:				
	10–19	20–29	30–39	40–49	50+
	0.66%	74.50%	16.78%	4.03%	4.03%
Q2	Please specify your gender:				
	Male		Female		
	53.70%		46.30%		
Q3	Are you involved in IT related jobs?				
	Yes		No		
	53.00%		47.00%		
Q4	On average, how many hours do you spend on watching videos on daily basis?				
	I do not watch videos	Less than 30 min	30 min - 1 hour	1 hour - 3 hours	More than 3 hours
	3.36%	36.24%	28.86%	26.17%	5.37%
Q5	On average, how many hours do you spend on listening to music on daily basis?				
	I do not listen to music	Less than 30 min	30 min - 1 hour	1 hour - 3 hours	More than 3 hours
	4.70%	29.53%	26.17%	28.19%	11.41%
Q6	How often do you listen to music while driving?				
	Never	Rarely	Occasionally	Frequently	Always
	4.70%	8.72%	14.09%	33.56%	38.93%

Q7	Please rate your preference for watching video using Smart Phones (e.g. iPhone)				
	1	2	3	4	5
	37.58%	24.16%	22.15%	12.75%	3.36%
Q8	Please rate your preference for watching video using Tablets (e.g. iPad)				
	1	2	3	4	5
	32.89%	9.40%	28.86%	19.45%	9.40%
Q9	Please rate your preference for watching video using Personal Computers				
	1	2	3	4	5
	4.70%	7.38%	14.09%	27.52%	46.30%
Q10	Please rate your preference for watching video using TVs				
	1	2	3	4	5
	8.72%	14.09%	14.77%	18.12%	44.30%
Q11	Please rate your preference for listening to music using Smart Phones / Music Players (e.g. iPhone / iPod)				
	1	2	3	4	5
	11.41%	6.71%	18.79%	26.85%	36.24%
Q12	Please rate your preference for listening to music using Tablets (e.g. iPad)				
	1	2	3	4	5
	33.55%	21.48%	23.49%	15.44%	6.04%
Q13	Please rate your preference for listening to music using Personal Computers				
	1	2	3	4	5
	8.72%	12.08%	21.48%	32.22%	25.50%
Q14	Please rate your preference for listening to music using Home Theater				
	1	2	3	4	5
	29.53%	17.45%	19.46%	12.75%	20.81%
Q15	Imagine yourself in the following scenario and rate how much it can enhance your experience. "You are watching a video on your mobile and you will be notified whenever there are other devices with better quality near you. You can select one of them and connect your mobile to it manually and continue watching the video on the new device."				
	1	2	3	4	5
	9.40%	12.75%	24.83%	34.23%	18.79%
Q16	Imagine yourself in the following scenario and rate how much it can enhance your experience. "You are watching a video on your mobile and the device with the best quality will be suggested to you. You can choose to switch to it or not."				
	1	2	3	4	5
	6.04%	10.07%	23.49%	26.85%	33.55%
Q17	Imagine yourself in the following scenario and rate how much it can enhance your experience. "You are watching a video on your mobile and whenever a device with better quality is available the video will be played on the better device automatically."				
	1	2	3	4	5
	13.42%	20.81%	17.45%	20.13%	28.19%

Answers

The first three questions (demographic questions) were discussed and reviewed in 3.2.1. As such, we continue with the three questions that belong to the second group: attitudes. We asked participants how much time they spend listening to music, watching videos and listening to music while driving. Participants had five options for each question. Figure 3.3 shows the distribution of the answers for the first two questions. As we can see in the Figure 3.3, participants had more distributed answer when it came to time spent watching videos. We will study the effect of these habits on the user satisfaction level of different devices in the following subsection. However, we should mention that people listen to music more than they watch videos, as evidenced by the mean of music listening (3.12) compared to that of watching videos (2.94) for watching videos (Table 3.6). Furthermore, as Table 3.6 demonstrates, most people always listen to music while they are driving.

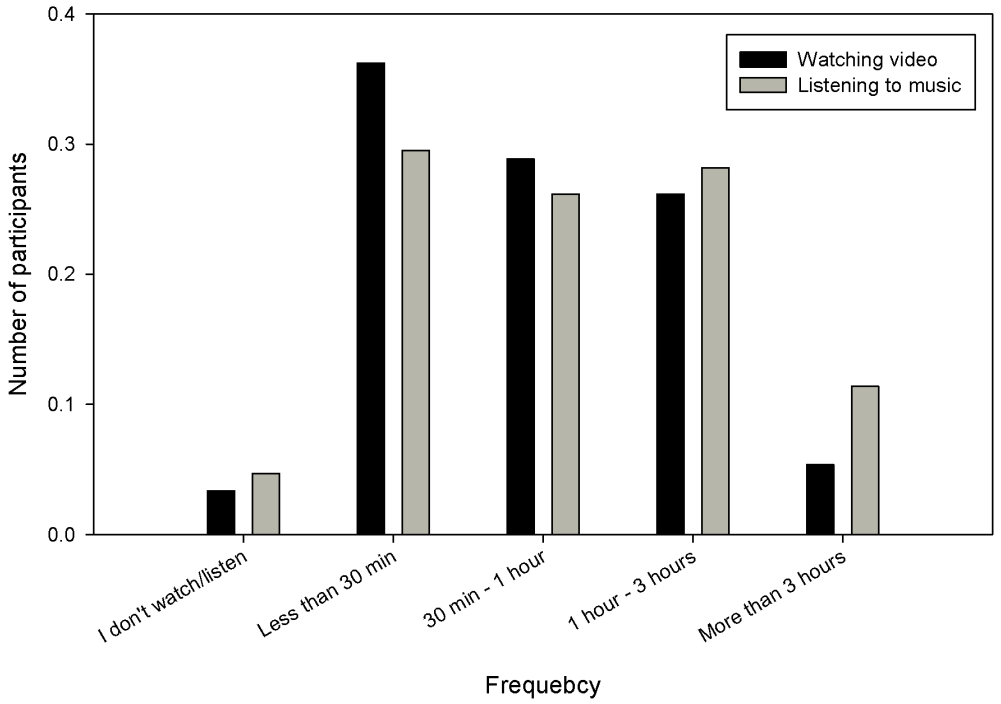


Figure 3.3: The distributions of answers to Q4 and Q5.

Table 3.6: Descriptive frequencies of answers to Q4, Q5, and Q6.

Index	Subject	Mean	Mode	Std. Deviation
Q4	Daily spending time on watching videos	2.94	2	0.988
Q5	Daily spending time on listening to music	3.12	2	1.102
Q6	Frequency of listening to music while driving	3.93	5	1.143

The main goal of this survey was to discover the preferences of our participants. Hence,

the majority of the questions in our questionnaire belonged to the third group. Music playback device preferences, video playback device preferences and desired media playback system were the three main subjects of the last group's questions. As Table 3.7 shows from the mean values of answers, people have different preferences for media playback devices. Also, their satisfaction level for each device differs from person to person. Personal computers are the favorite devices for watching videos, followed by TVs. It is a completely different story for music players. Smart phones and music players are at the top of the list and personal computers are on the second place.

What's more, the questions related to the desired system allowed for an interesting discovery. They showed that people prefer to be informed about the best available alternative devices while they want to be in control of accepting or rejecting the media redirection. Moreover, as the mean values of the last three questions are different, we should also pay attention to their mode. As Table 3.7 shows, the mean value of Q15 is bigger than Q17 but the mode of Q17 is 5 while it is 4 for Q15. This shows that participants are more distributed for Q17. The next section will demonstrate the correlation between the participants' preferences and their answers to demographic and attitude questions.

Table 3.7: Descriptive frequencies of questions Q7 to Q17.

Index	Subject	Mean	Mode	Std. Deviation
Q7	Video on Smartphone	2.20	1	1.174
Q8	Video on Tablet	2.63	1	1.362
Q9	Video on PC	4.03	5	1.153
Q10	Video on TV	3.75	5	1.375
Q11	Music on Smartphone	3.70	5	1.329
Q12	Music on Tablet	2.39	1	1.261
Q13	Music on PC	3.54	4	1.239
Q14	Music on Home theater	2.78	1	1.511
Q15	Manual System	3.40	4	1.202
Q16	Semi-automatic System	3.72	5	1.203
Q17	Automatic System	3.29	5	1.416

Statistical Analysis

This subsection presents the statistical analysis of the answers obtained. There are three independent variables in the questionnaire: age, gender, and profession. We believed that those variables were the likeliest to affect the users' preferences, hence we studied their effects on user satisfaction levels. There are also three other variables that are dependent on the first group the attitudes of our participants but they can be considered as independent variables when studying their effects. This subsection will demonstrate the analysis of answers based on the aforementioned variables.

- **Age:** The first independent variable is age range. Participants could answer this question with five different options. As we discussed earlier, the average age of the

participants in the conducted survey was around 28 years. Due to this, we decided to regroup answers into two categories - Young ($age \leq 30$) and Old ($age \gneq 30$) - before we studied the effect of age range. Then, we examined the t-Test for equality of means for these two categories. Table 3.8 shows the results of this test.

Table 3.8: The t-test for equality of means of different age groups.

	Levene's Test for Equality of Var.		t-Test for Equality of Means		Means	
	F	Sig.	t	Sig.	Young	Old
Q7	0.000	0.993	-0.557	0.580	2.23	2.11
Q8	0.683	0.410	-1.149	0.255	2.71	2.41
Q9	15.885	0.000	-3.265	0.001	4.21	3.51
Q10	1.587	0.210	0.025	0.980	3.75	3.76
Q11	5.532	0.020	-1.117	0.266	3.77	3.49
Q12	0.117	0.733	-2.419	0.018	2.53	1.97
Q13	5.734	0.018	-3.470	0.001	3.73	2.95
Q14	1.499	0.223	0.142	0.887	2.77	2.81
Q15	3.281	0.72	-1.406	0.166	3.49	3.14
Q16	2.125	0.147	-0.660	0.512	3.76	3.59
Q17	0.966	0.327	1.330	0.189	3.20	3.57

Table 3.8 shows that age can affect the user satisfaction level for different devices. As the Sig. values are less than 0.05 for Q9, Q12, and Q13, we can claim that age range can affect the user preferences and it, thusly, should be considered a variable. The mean value of answers to the questions posed shows that people lose interest in listening to music on their PC or in watching videos on tablets or PCs as their age increases. In addition, table 3.8 shows that age has a slight effect on Q15 and Q17, which is to say that older people prefer automatic systems to manual system.

- **Gender:** When we were studying the effects of gender on user preferences, we were expecting to see clear and promising relations but statistical tests rejected our assumptions; we could not find a robust and direct relation between them. Then, we studied the effects of gender on users' attitudes. Table 3.9 shows the results of this study.

Table 3.9: The t-Test for equality of means of attitudes grouped by gender.

	Levene's Test for Equality of Var.		t-Test for Equality of Means		Means	
	F	Sig.	t	Sig.	Male	Female
Q4	4.270	0.041	0.636	0.526	2.99	2.88
Q5	0.932	0.336	-1.904	0.060	2.96	3.30
Q6	0.997	0.320	-1.243	0.213	3.82	4.06

As we can see in table 3.9, there is a trivial relation between gender and user attitudes that allows us to claim that males watch more videos than females. On the other hand, females listen to more music more than males whether or not they are driving.

Later, we will see the influence of attitudes on user preferences. To summarize, we can say that gender has an indirect effect on user satisfaction.

- **Profession:** The last independent variable is the users' profession. We were expecting to see that the users' backgrounds and knowledge could affect the preferences. Table 3.10 shows the result of t-Test for equality of means that confirms the expected results.

Table 3.10: The t-Test for equality of means based on job type.

	Levene's Test for Equality of Var.		t-Test for Equality of Means		Means	
	F	Sig.	t	Sig.	No	Yes
Q7	4.276	0.040	0.126	0.900	2.21	2.19
Q8	3.197	0.076	-2.209	0.29	2.37	2.86
Q9	20.908	0.000	-3.285	0.001	3.71	4.32
Q10	5.240	0.023	-2.128	0.035	3.50	3.97
Q11	0.724	0.396	0.385	0.701	3.74	3.66
Q12	9.083	0.003	0.227	0.821	2.41	2.37
Q13	10.689	0.001	-2.089	0.038	3.31	3.73
Q14	1.473	0.227	-0.810	0.419	2.67	2.87
Q15	2.143	0.145	-0.025	0.980	3.40	3.41
Q16	0.001	0.977	0.235	0.814	3.74	3.70
Q17	0.258	0.612	0.903	0.210	3.40	3.19

As table 3.10 shows, Q9, Q10, and Q13 are strongly related to the job type since they have small sigma values. This means that users who work in IT industries like to watch videos on their PCs and TVs more than users who do not. In addition to this, IT workers like using smartphones/music players as well as PCs to an almost equal degree to listen to music while non-IT workers prefer smartphones/music players to any other music-playing devices. In conclusion, we can declare that job type can be considered another influential variable on user preferences determination.

- **Attitudes:** As we mentioned earlier, we believed that users' attitudes and habits could affect their preferences and satisfaction level. Since we grouped user attitudes into five categories, we could not perform the t-test to examine the possible relationships between these two variables. In such situations, the One-way ANOVA test is commonly used. Table 3.11 shows the results of this test for Q7 to Q14. Q7 to Q10 were studied based on video-watching habits and Q11 to Q14 were examined according to music listening habits.

  There are several relations in this table. As far as we can see, there is a direct and clear relation between the frequency at which a user watches videos and their satisfaction level with TVs. Also, the frequency at which they listen to music affects their satisfaction level with PCs and smartphones/music players. As a result, we can say that users' habits and attitudes can influence the level of satisfaction they experience with various types of media playing devices as well as their preference for

Table 3.11: The One-way ANOVA test results for device preferences based on the user attitudes.

	One-way ANOVA test		Means				
	F	Sig.	G1	G2	G3	G4	G5
Q7	2.613	0.038	1	2.06	2.16	2.56	2.38
Q8	2.110	0.083	1	2.63	2.60	2.82	2.88
Q9	2.924	0.023	2.40	4.06	4.14	4.03	4.38
Q10	1.043	0.387	3	3.65	3.65	4.03	4.12
Q11	3.240	0.014	2.71	3.39	4.10	3.62	4.18
Q12	1.706	0.152	1.71	2.36	2.51	2.19	2.94
Q13	4.764	0.001	2.29	3.23	3.49	3.86	4.18
Q14	1.059	0.379	1.86	2.98	2.59	2.83	2.78

one such device over another. Hence, user habits and attitudes should most definitely be counted as variables in any system that wants to determine user preferences.

Chapter 4

ProGes

The IoT has become a hot research topic in recent years. Consequently, several studies have tried to develop methods that are suitable for interacting with the IoT. These studies selected various approaches. Some studies reviewed interactions among the objects within the IoT. For example, Kortuem et al. studied the different interactions between smart objects in the IoT [30]. The second group of studies focused on the interactions between humans and objects such as [7, 9, 20]. Another group of studies explored the user-centered interactions in the IoT to elicit guidelines and propose generic frameworks [10, 31]. However, none of these researches have specifically focused on user and multimedia objects interaction over the IoT.

This chapter presents the proposed method, ProGes, which is a multimodal interaction method for controlling multimedia devices over the IoT. ProGes covers two types of interactions: Human-Device interaction and Device-Device interaction. In addition to that, ProGes is designed considering the findings of Kranz et al.: invisibility dilemma, implicit vs. explicit interaction, context dependence, interaction and multimodality [31]. First, we present the overall architecture of ProGes. Then, detailed specifications of ProGes are given.

4.1 System Architecture

ProGes uses proxemic and gesture interaction to provide a natural interaction environment for communication between multimedia devices and humans. Moreover, ProGes collaborates with local services such as tracking and gesture recognition engines and online services like cloud-based media streaming. High-level architecture of ProGes is presented in Figure 4.1.

ProGes consists of four units: Interaction Proxy Unit (IPU), Proximity Interaction Unit (PIU), Gesture Interaction Unit (GIU) or Gesture Control Unit (GCU), and Media Redirection Unit (MRU). The Interaction Proxy Unit has two main responsibilities. First, it acts as the external interface of ProGes. Accordingly, devices and users in the IoT interact with ProGes through calling the IPU's service (e.g. registration, update location, etc.).

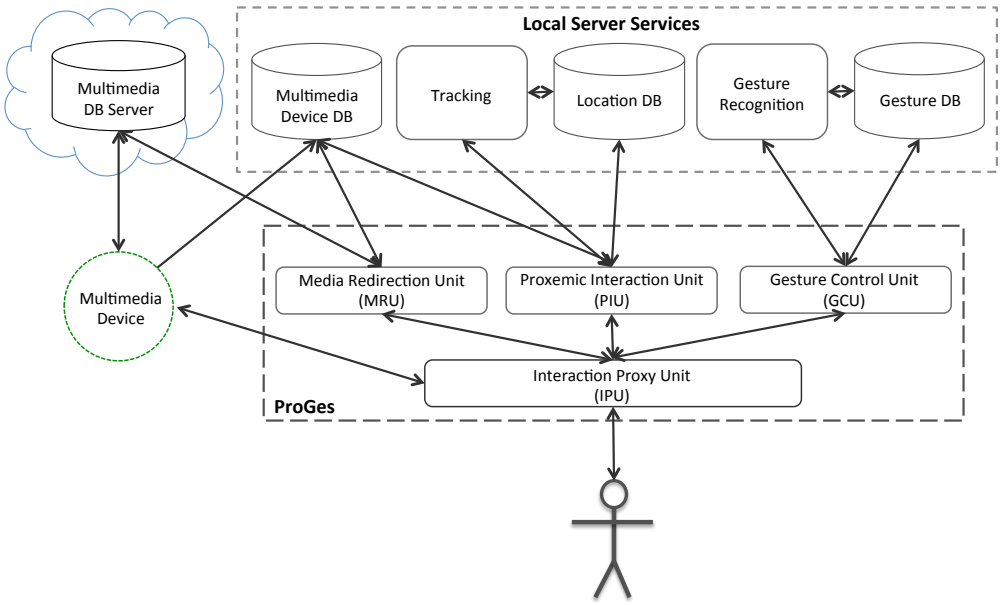


Figure 4.1: The abstract architecture of ProGes and its units. This figure presents the existing connections between 4 units of ProGes and local and cloud-based services. Moreover, we can see that users and devices can only communicate with ProGes through the Interaction Proxy Unit.

Second, it has the responsibility for coordinating ProGes's units and their communications. Since Interaction Proxy only has access to inputs and outputs of ProGes, it delivers messages to different units and keeps them independent. The Proxemic Interaction Unit finds the possible proxemic interactions and potential media redirections, and informs the Interaction Proxy about them afterwards. The Gesture Control Unit receives video streams through the IPU, followed by recognition of gestures and returning the detected control commands to the IPU. The Media Redirection Unit handles the records of users and devices all the time. It also preprocesses media streams and prepares them for devices according to devices' specifications. Furthermore, the current trend is that a cloud-base database keeps the shared resources including multimedia data. Hence, we assumed a cloud-base Multimedia Database and streaming server in our design which is in accordance with the current technological trends.

There are three databases in our architecture: Multimedia Device database which is responsible of keeping device specifications, Location database that stores the location of devices and users, and Gesture database that acts as a dictionary for gestures and control commands. Moreover, two local services are working closely with our system: a Tracking service that uses WiFi signals to find the location of users and devices and a Gesture Recognition service that receives the normalized data and detects the performed gestures. The following sections present detailed description of each component with the help of two use case scenarios.

4.1.1 Use-case Scenarios

We present two use-case scenarios in this section. The first scenario explains an interaction between two multimedia devices while the second one presents an example of human-device interaction. We use SDEdit ([62]) to draw sequence diagrams in order to present our use-case scenarios.

Device-Device Interaction Scenario

The first scenario shows the proxemic interaction between two devices: device1 and device2. Figure 4.2 shows this scenario which presents the device-device media redirection interaction.

When device1 and device2 connect to the network, they register themselves in the Multimedia Device Database. Next, device1 and device2 create new threads to send their location information to the Interaction Proxy Unit (IPU). The IPU processes the information and finds the current location of them in collaboration with the PIU and Tracking service. Then, the current location of device1 and device2 is updated in the Location Database. Separated threads execute this process repeatedly for device1 and device2 in the loops. These loops stop only when a device sends a stop request and deletes itself from the Multimedia Device Database.

Next, device1 sends a request for media content to the Media Streaming Server, which is located on the cloud. Once device1 receives the media stream, it notifies the IPU that it is ready to participate in a proxemic interaction. When the Interaction Proxy receives this notification, it notifies the PIU to create a new thread and look for a possible proxemic interaction. In this thread, the PIU first retrieves the current location of all the registered devices and extracts a list from it. This list only includes devices that are in the proximity of device1 (device2 is the only device on the list in this scenario). Then, PIU collects the devices' specifications, media content properties, and user's profile information. Combining these 3 groups of data, PIU calculates a score for each device and finds the device with the highest score (device2). As a result, it recommends device2 to the IPU. Next, the IPU sends a preparation signal to Media Redirection Unit (MRU) for device2. Once the IPU receives the ready confirmation from MRU, it sends a notification to device1 and informs the user about the possible redirection. If user rejects this recommendation, the IPU sends a cancelation signal to MRU and starts to find another device. But, if user accepts this recommendation, the IPU sends a redirection command to MRU. Then, MRU prepares the stream for device2 and it starts playing the media content stream form the server. Afterward, the IPU sends a stop command to device1 after device2 starts playing. In the end, device1 stops the stream and disconnects from the Multimedia Stream Server.

In conclusion, as this scenario described, the Interaction Proxy play the roll of the main coordinator in the ProGes. Proxemic interactions are not possible without a tight collaboration between the Interaction Proxy and other units.

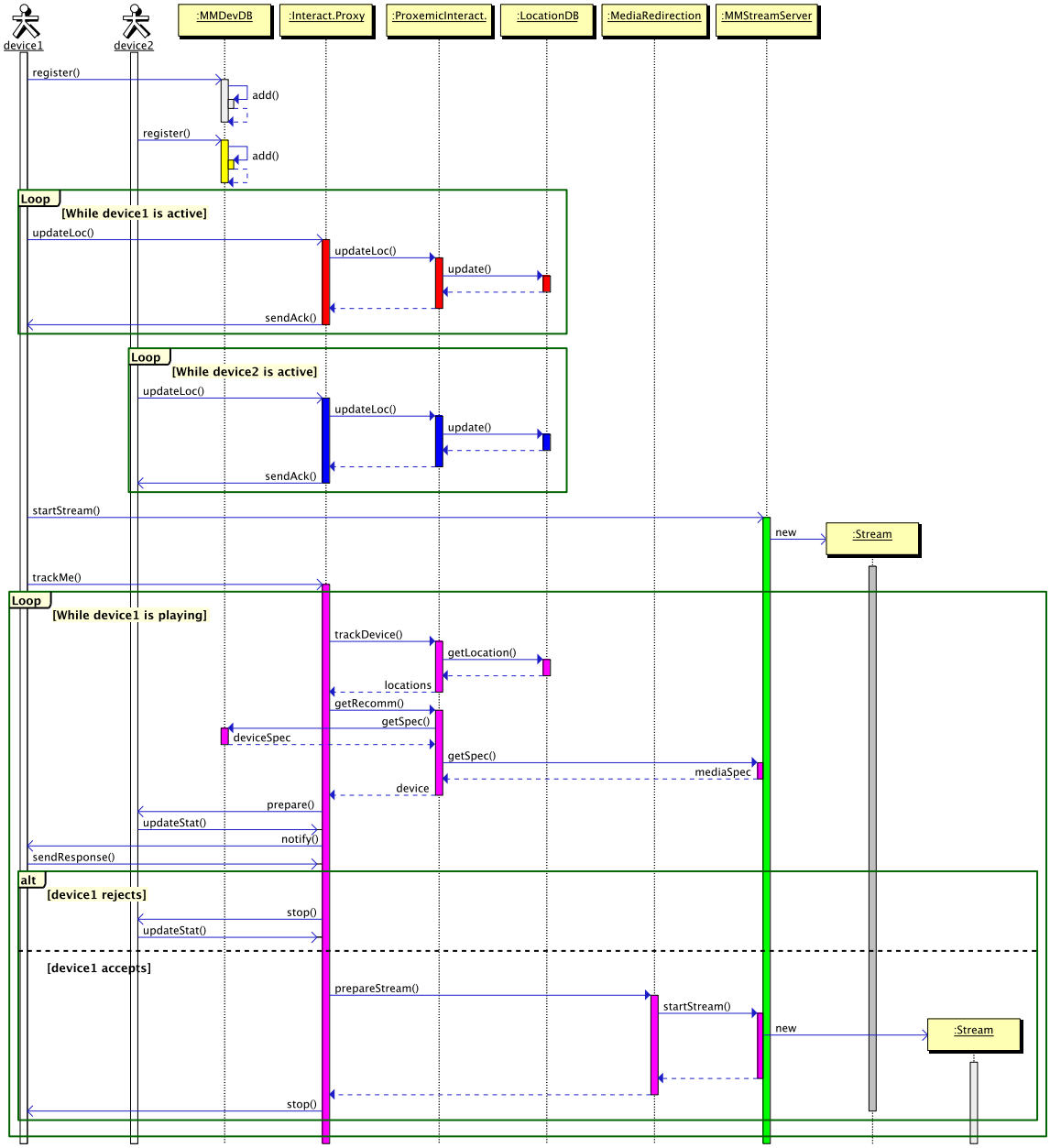


Figure 4.2: An example scenario of device-device interaction within ProGes. It shows that when two devices get close to each other, ProGes calculates scores for those devices and considers the possibility of multimedia redirection between them.

Human-Device Interaction Scenario

Figure 4.3 shows the second scenario. This scenario presents a possible simplified Human-Device interaction. It only has one user and one device. The main goal of this scenario is clarifying the functionality and advantages of the combination of proxemic and gesture interaction.

This scenario starts, like all other possible scenarios, by device registration. When device connects to the network, it registers itself in the Multimedia Device Database. Subsequently, it creates a new thread to send its location information to the IPU in an infinite loop. Once the IPU receives the location information, it uses the Tracking service to find the latest location of device and updates it in the Location Database. At the same time, the IPU tracks the user. It uses the Tracking service to follow his or her movements. If there is a change on the current location, it updates it in the Location Database. The IPU tracks both device and user until it receives a termination request from them.

When the location of user is updated in the database, the IPU sends an engagement check signal to the PIU. The PIU examines whether the user is in a pre-defined proximity of a gesture-enabled device or not. In this scenario, user's new location is within the proximity of device, which is a gesture-enabled one. Once the IPU receives a true value on the engagement flag from the Proxemic Interaction unit, it sends start tracking command to the Gesture Control Unit. Then, the GCU creates a new thread to process the video of user and detects any preformed gestures. This thread asks the IPU to capture and provide a live video stream feed to the GCU. When the GCU receives the video, it extracts the pre-defined features from it and sends them to the Gesture Recognition Unit. Subsequently, the Gesture Recognition Unit starts detecting gestures using an assist from the Gesture Database. Once a gesture is detected, its code is used to retrieve the corresponding command for the device. Later, the command is given to the GCU. The GCU hands in the command to the IPU where the command is passed to the device for execution. The main advantage of this combination is detecting the active user and among all humans in the line of sight of the gesture-enabled device and executing his/her commands.

To conclude, this scenario emphasizes a main advantage of ProGes, which is the combination of proxemic and gesture interaction. To do that, it shows how the proxemic interaction complements the gesture interaction by providing an intuitive engagement mechanism.

4.2 Interaction Proxy Unit

The Interaction Proxy is the backbone of ProGes. It has two major responsibilities. First, it acts as the public interface of the system. Therefore, all users and devices are connecting to it in order to communicate with ProGes. It receives the location information data from devices and also it collects the captured video streams from gesture-enabled devices. Then, it processes the data and redirects it to the corresponding unit. For example, when it receives the location information from a device, it sends that data to the Proxemic Interaction Unit. The PIU will update the location of that device and detect possible proxemic interaction.

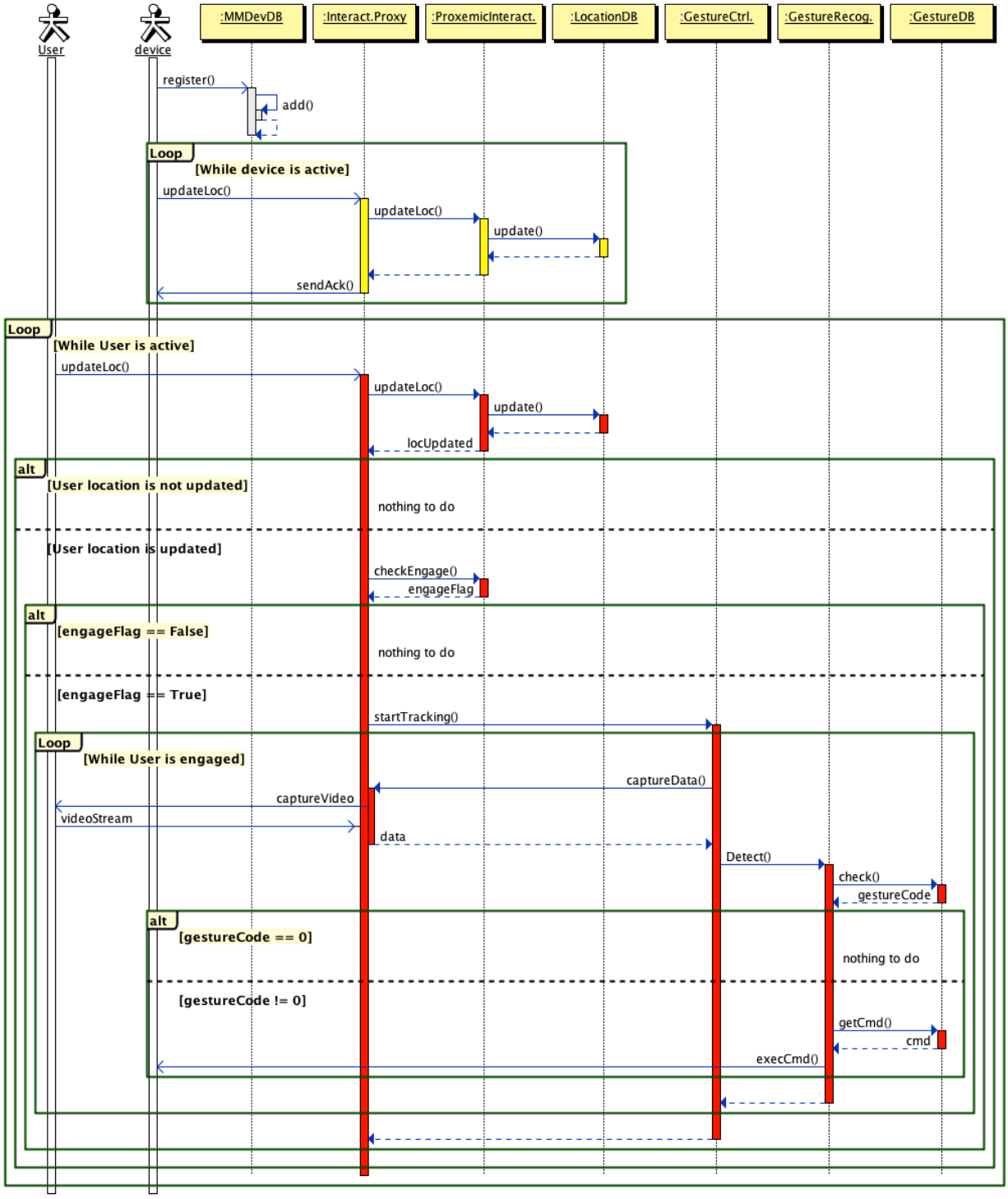


Figure 4.3: An example scenario of human-device interaction within ProGes. It shows that when a user is engaged with a gesture-enabled device and enters into its perimeter, ProGes enables the gesture recognition and tracks that user among other people using the proxemic information.

The IPU also receives the requests and generated commands by its subsidiaries and handles them accordingly. A good example is when the PIU detects a possible proxemic media redirection. It notifies the IPU and IPU informs the Media Redirection Unit. The MRU handles the rest. Engaging a user in a gesture control mechanism is similar to that while IPU informs the GIU instead of the MRU.

4.3 Proxemic Interaction Unit

Proxemic Interaction Unit is the central part of the system as it is the decision maker unit. Figure 4.4 presents the abstract algorithm of PIU. As we can see in the figure, the algorithm begins by receiving the location information of the devices. PIU uses the tracking engine to find the locations and update it on the location database. Next, PIU checks for possible proxemic interaction. Then PIU calculates a score between 0 and 1 for each device based on content properties, user preferences, and device capabilities. Consequently, the algorithm finds the device with maximum score. If it is not the same device as the one in use, it notifies the user through IPU and recommends a media redirection. If user accepts the recommendation, IPU asks the MRU to handle this process. After each interaction, PIU updates the scoring coefficients to learn user's preferences adaptively.

The device scoring and recommendation step in the aforementioned algorithm is the most important step since it can assist users to engage with a larger number of devices in the IoT. To recommend the most appropriate device in the given scenario, we should find and imply different variables in the scoring mechanism. Therefore, we start by introducing these variables and then present the scoring mechanism. There are three groups involved in the proposed solution: devices, users, and multimedia content. Below we explain these groups in order to find effective variables for the scoring mechanism.

4.3.1 Devices

Devices are the first and most important objects in our system. Since the proposed system provides an environment for proxemic interactions, we begin by explaining device's proximity behavior. Hall defined four perimeters around each person: intimate space, personal space, social space, and public space (Table 4.1) [21]. We are using same categorization for devices according to their effective quality in each space: intimate devices, personal devices, social devices, and public devices. For example, smartphones have small screens, so they are usually used by individuals separately. Moreover, they have the best visual quality when they are close to the user (e.g. 20 cm which is in the intimate space). Users can still see the smartphone screen when they are farther away, but their effective quality decreases. Similarly, users enjoy big screen TVs most when they are at an acceptable distance, i.e., 3.5 m to 6 m. Hence, they are placed in the category of public devices. Let $T(x)$ be a function that returns an integer between 0 to 3 depending on the type of device x. Table 4.1 provides examples of all 4 groups of device spaces and function $T()$ value for each device type. We will use function $T()$ later while calculating score for each device.

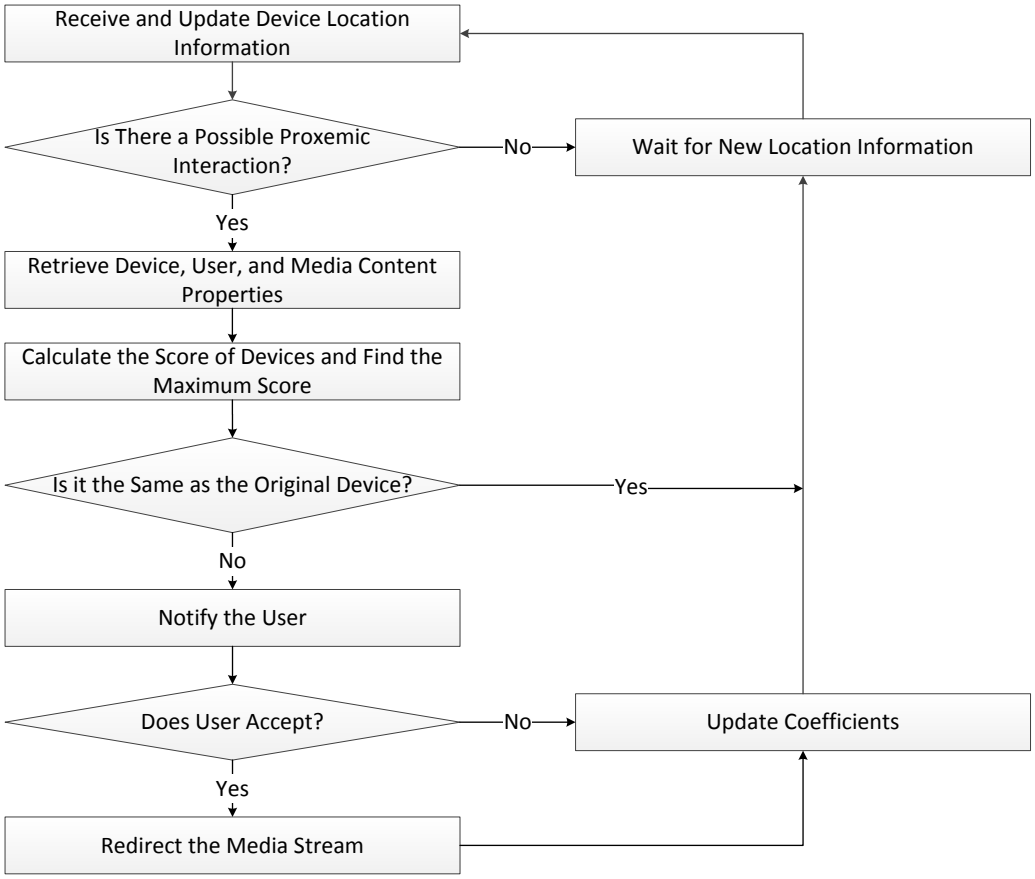


Figure 4.4: The abstract algorithm of the Proxemic Interaction Unit.

Table 4.1: Hall's personal space definitions and device examples for each space.

Space Name	Space Area	Device Example	$T()$
Intimate space	$distance \leq 0.45m$	Smartphones	0
Personal space	$0.45m \lneq distance \leq 1.2m$	Tablets, Laptops	1
Social space	$1.2m \lneq distance \leq 3.6m$	PCs, Digital displays	2
Public space	$3.6m \lneq distance \leq 7.6m$	TVs, Home-stereos	3

4.3.2 Users

Since we are designing a UI, users should be the most important element in our design. As we explained earlier in section 3.2, we conducted a user survey to elicit user's preferences on the multimedia devices. Results of this user survey revealed some interesting points. First, we could not find any correlation between the gender and device preferences. However, when we grouped our participants into young (≤ 30 $years$) and old (> 30 $years$) users, we found that old participants are more interested in using TVs and PCs than tablets and smartphones. In addition to that, profession had an effect on the device preferences. Participants who were involved in jobs that need a high level of IT knowledge were keener to use PCs and TVs. We also found that frequent users (who spend up to 1 hour per day) prefer TVs and PCs more than non-frequent users (who spend more than 1 hour per day).

To summarize, based on the results of this user survey, we were convinced to include three characteristics of the engaged user, u, in our scoring mechanism: participant's age (U^a), profession type (U^p), and multimedia usage habits (U^h). We used the user ratings to initialize preference coefficients corresponding to these characteristics, which are updated over time according to user interactions.

4.3.3 Multimedia Contents

Multimedia content properties can influence the user's decision regarding the playback device. For example, a video content cannot be streamed to a home stereo, which only supports audio inputs. So, the type of the content should be considered. In addition, there is a user study that shows the length of multimedia content can affect the user's choice for playback device [37]. It categorizes videos based on their length: very short videos (up to 2 minutes, e.g. social network clips), short videos (up to 7 minutes, e.g. music videos), medium videos (up to 22 minutes, e.g. soap operas and animations), long videos (up to 45 minutes, e.g. TV series), and very long videos (longer than 45 minutes). Results of this study indicates that the length of the first half of videos that are played on the smartphones over the WiFi connection is around 50 seconds while it is almost 100 seconds for tablets. This is due to several factors such as screen dimensions and resolutions, battery capacity, etc. Hence, we decided to use three properties of the given multimedia content, c, in our scoring mechanism: audio of content (C^a), video of content (C^v), and duration of the content (C^d). We have grouped long and very long videos together since there is no limitation for the content's length, as a result, C^d can take four values depending on the length: 0 - very short, 1 - short, 2 - medium, and 4 - long.

4.3.4 Scoring Mechanism

When the PIU finds out that there is a possible proxemic interaction, it starts the scoring mechanism. Let $D = \{D_i | 1 \leq i \leq n\}$ be the set of available devices. For the given content c and a user u, the score of i^{th} device, s_i, is calculated as follows:

$$s_i = f_{ic} \times m_{ic} \times h_{iu} \times a_{iu} \times p_{iu} \times d_{iu} \tag{4.1}$$

where f_{ic} is a flag that indicates capability of device D_i to play content c, m_{ic} is device appropriateness coefficient for the given content, h_{iu}, a_{iu}, and p_{iu} are user's habit, age, and profession factors for the device D_i. The last element is d_{iu} is the device effectiveness for the distance between device D_i and user u. The value of all of elements is defined between 0 and 1. We have chosen to multiply individual factor to obtain s_i because for optimal results we want all factor to be unity; on the other hand, if even one factor is zero, the device is useless. For example, if a device cannot play the given content, the value of other elements are not important at all and that device should get 0 score. In the following text, we explain each element in details.

Playback Capability

This element represents the playback capability of D_i for the given multimedia content c. It is a binary variable, so it is either 0 (i.e. i^{th} device cannot play c) or 1 (i.e. i^{th} device can play c). Equation 4.2 shows the Boolean expression that is used to compute f_{ic}:

$$f_{ic} = (D_i^a \odot C_c^a) \wedge (D_i^v \odot C_c^v) \tag{4.2}$$

where D_i^a and D_i^v logical variables which are true if device D_i is capable of playing audio and video respectively, otherwise false. Similarly, C_c^a and C_c^v are true if the given multimedia content c has audio and video playback requirements, otherwise false.

Device Appropriateness

As we explained in subsection 4.3.3, the length of given content (C_c^d) can affect the user's playback device preferences. The value of m_{ic} is selected from the 4×4 matrix (M), which its rows represent the length of the given content and its columns show the device space types:

$$m_{ic} = M(C_c^d, T(D_i)) \tag{4.3}$$

The first row of M is dedicated to the content with $length \leq 2\ min$, second row for the content with length between 2 min and 7 min, third row for the content with between length 7 min and 22 min, and the last row covers the rest ($length > 22\ min$). The columns follow this order: intimate devices, personal devices, social devices, and public devices. The same order is followed in the other defined matrices that have the device space type as column indicator. To define the thresholds and matrix entries' initial values, we used the results of Li et al.'s study [37]. They conducted a broad study in China on over 3 million users for 2 weeks. The defined matrix is as follows:

$$M = \begin{bmatrix} 1 & 0.8 & 0.6 & 0.4 \\ 0.8 & 1 & 0.8 & 0.6 \\ 0.6 & 0.8 & 1 & 0.8 \\ 0.4 & 0.6 & 0.8 & 1 \end{bmatrix} \tag{4.4}$$

Habit, Age, and Profession

To determine the factors of user's multimedia usage history (h_{iu}), age (a_{iu}), and profession (p_{iu}); we define three binary functions as follows:

$$\begin{aligned} U_u^h &= \begin{cases} 0 & \text{if } u \text{ is a frequent user} \\ 1 & \text{if } u \text{ is a non-frequent user} \end{cases} \\ U_u^a &= \begin{cases} 0 & \text{if age of } u \text{ is less than or equal to 30} \\ 1 & \text{if age } u \text{ is greater than 30} \end{cases} \\ U_u^p &= \begin{cases} 0 & \text{if } u \text{ does a high IT knowledge required job} \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{4.5}$$

The factor h_{iu} depends on the given user's multimedia usage habits. As we explained earlier (section 3.2), user's multimedia playing time can influence the preferred playback device. The habit factor h_{iu} is calculated as follows:

$$h_{iu} = H(U_u^h, T(D_i)) \tag{4.6}$$

where H is defined as a 2×4 matrix where rows represent usage frequency (frequent, non-frequent) and columns represent device space type. The age and profession factors are also calculated in the similar way as follows:

$$a_{iu} = A(U_u^a, T(D_i)) \tag{4.7}$$

$$p_{iu} = P(U_u^p, T(D_i)) \tag{4.8}$$

where A and P are again 2×4 matrices. The first row of A keeps a_{iu} values for users who are younger ($U_a \leq 30\ years$) and the second row presents a_{iu} values for the older users ($30\ years \gneq U_a$). Similarly, first row of P represents users who work in a high IT knowledge-required job and the second row presents rest of the users. The initial values of matrix elements of H, A, and P are determined based on user rating in the survey. Let $Y = \{Y_j | 1 \leq j \leq 8\}$ be the 8 mean values in each categories and $W = \{W_j | 1 \leq j \leq 8\}$ be their corresponding initial values in H, A, or P. Then, W_j is calculated as follows:

$$W_j = \lceil \frac{Y_j}{max(Y)} \rceil \tag{4.9}$$

where the result of the division is round up to the closest decimal point. The calculated initial values for these three matrices are as follows:

$$H = \begin{bmatrix} 0.8 & 0.7 & 1 & 0.9 \\ 0.8 & 0.8 & 0.9 & 0.9 \end{bmatrix} \tag{4.10a}$$

$$A = \begin{bmatrix} 0.8 & 0.7 & 1 & 0.9 \\ 0.8 & 0.6 & 0.9 & 0.9 \end{bmatrix} \tag{4.10b}$$

$$P = \begin{bmatrix} 0.8 & 0.7 & 1 & 0.9 \\ 0.8 & 0.6 & 0.9 & 0.8 \end{bmatrix} \tag{4.10c}$$

Distance Effectiveness

The distance effectiveness d_{iu} is calculated using the distance between the given user u and device D_i. Let x_{iu} be the distance between user and device. We defined a unique device space effectiveness function for each type of device (where type is defined in terms of space). Figure 4.5 shows the plot of the defined functions. These functions are extracted from the observations on device manuals and user's actions. For example, a big screen TV is mostly effective when $3.5\ m \leq x_{iu} \leq 6\ m$. On the other hand, it is not convenient for the user to watch the TV when the distance is around $0.5\ m$. You can see this behavior on the Public function in Figure 4.5. Similarly, user cannot clearly see a smartphone screen when too close ($x_{iu} \approx 0\ m$). However, the smartphones are very effective when x_{iu} is around $0.25\ m$. Afterwards, when x_{iu} increases, their effectiveness (d_{iu}) decreases due to their small screen size. Generally, we can argue that the effectiveness function is asymmetric around its peak.

Finally, we could fit Weibull II with four parameters on the observed curves in order to define effectiveness function. It is restricted between 0 and 1. Equation 4.11 shows the general form of this function where x_{iu} is the distance between the user and device; and x_0, a, b, and c are four parameters, which are different for each device spaces.

$$d_{iu} = a \times (\frac{c-1}{c})^{\frac{1-c}{c}} \times \left| \frac{x_{iu} - x_0}{b} + (\frac{c-1}{c})^{\frac{1}{c}} \right|^{c-1} \times e^{-\left| \frac{x_{iu}-x_0}{b} + (\frac{c-1}{c})^{\frac{1}{c}} \right|^c} + \frac{c-1}{c} \tag{4.11}$$

4.3.5 Adaptation Mechanism

The proposed solution considers different variables to calculate the score for each device (Equation 4.1) in order to find the best possible device for media redirection. Although we used the results of different surveys in order to define the values of the coefficients in the scoring mechanism, each person may have different preferences. Therefore, we designed an adaptation mechanism in the proposed solution, which updates the coefficient values based

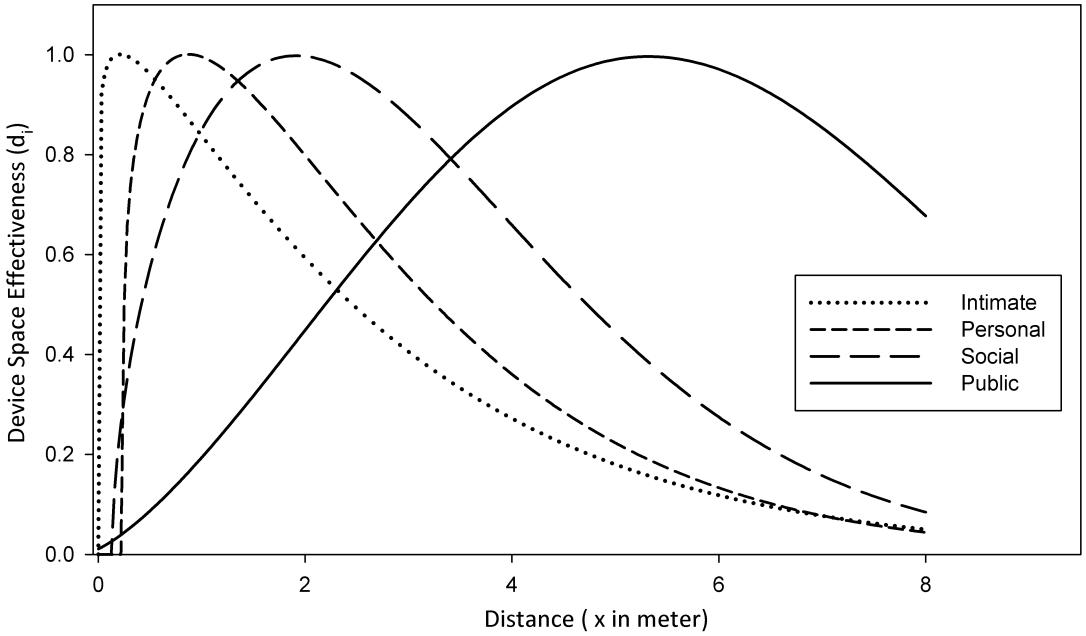


Figure 4.5: The plot of 4 different distance functions.

on the responses of individuals after each interaction. The PIU selects the corresponding values for coefficients and calculates the scores. For example, imagine PIU uses user's age coefficient a_{iu} to calculate the score for device D_i, and D_i is recommended to the user since it has the highest score. So, when the user responds to the recommendation, the value of a_{iu} is updated using the following equation:

$$a_{iu,new} = a_{iu,old} \times e^{0.001 \times r_{iu} \times (1 + \frac{n_a}{n_o + 1})} \tag{4.12}$$

where $a_{iu,new}$ is the new coefficient's value, $a_{iu,old}$ is the old coefficient, r_{iu} is the effect of user's response to recommended device (D_i) which we will explain later (Equation 4.13), n_a is number of accepted recommendations for this type of device space ($T(D_i)$), and n_o is the overall number of recommendations for this type of device space ($T(D_i)$).

User Response Effect

To consider the effect of user's response on the adaptation mechanism, we defined factor r_{iu}, which depends on the user's response to the device recommendation. As we explained earlier, user's response to recommendation can influence the new values of used coefficients. The response factor r_{iu} is calculated as follows:

$$r_{iu} = \begin{cases} -1 & \text{if } u \text{ rejects the recommendation} \\ +1 & \text{if } u \text{ accepts the recommendation} \end{cases} \tag{4.13}$$

4.4 Media Redirection Unit

The Media Redirection Unit (MRU) has two main responsibilities. First, it keeps the records of users, devices, and multimedia streams in a database. This data repository is a great asset for finding the required information for proxemic interactions and handling multimedia redirections.

The second task of MRU is more important. It should preprocess the media stream and prepare it for target devices. Imagine the user is watching an HD movie on a big screen TV in the living room. Suddenly, s/he decides to leave the room. Therefore, the video should be redirected to the smartphone of the user. However, the smartphone cannot play the HD video. Thus, the MRU reduces the quality of the video to make it proper for the smartphone. This has few benefits among them we describe two: first, it save the bandwidth on the network. Second, it can reduce the process on the smartphone which can save more energy and increase the battery life of the device.

4.5 Gesture Control Unit

The Gesture Control Unit is responsible for providing the gesture interaction experience to the users. It fulfills this responsibility by finding the engaged user first, and then detecting the user's gestures. Next, using the Gesture Database, it maps the detected gesture to a control commands that are executable by the involved device. The GCU and the gesture recognition engine works together in order to implement the elicited gesture vocabulary and providing the natural user interface to the users. Figure 4.6 shows the step-by-step level flowchart of the this unit.

The GCU is involved in the ProGes when the IPU sends a start-tracking signal. When a user moves and enters into a perimeter of a device, which is gesture-enabled, the PIU notifies the IPU. Then, IPU sends this signal to the GCU. Next, the GCU requests the captured video data stream from the IPU. The IPU sends the captured data stream to the GCU. If there is more than one person in the captured data, the GCU uses the proximity information to detect the involved users. Consequently, the GCU prepares the required information by extracting features from the captured data and gives them to the gesture recognition service. If gesture recognition service detects a gesture, it returns the gesture code to the GCU. The uses the Gesture database to translate the gesture code to the device's control command. The control command is sent back to the IPU in the executable format. The IPU sends the command to the corresponding objects for execution in the end.

We should emphasize that the implementation of gesture recognition service is completely independent from the functionality of the ProGes. So, gesture recognition can have different implementations based on the variety of techniques. As we mentioned earlier in section 2.2.1, we adapted a DTW algorithm to implement the gesture recognition service for the proposed gesture vocabulary.

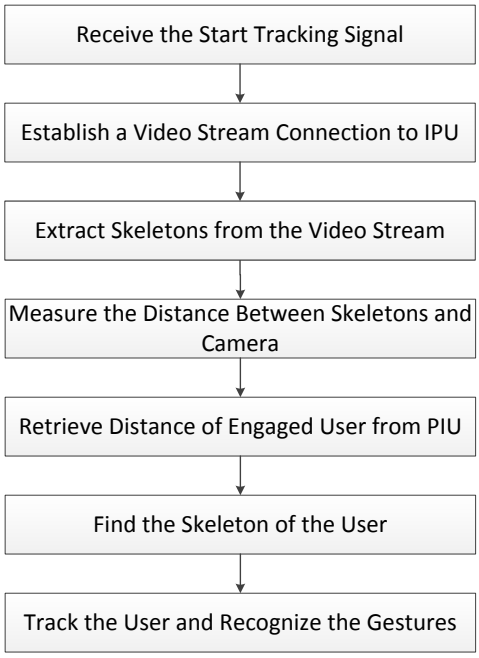


Figure 4.6: The step-by-step flowchart of the gesture control unit.

Chapter 5

Evaluation and Discussion

This chapter is dedicated to evaluation and discussion. In the beginning, a two-step evaluation on gesture vocabulary is given. The first step is an external user agreement which is presented to validate the proposed vocabulary. Then, it gives the description and results of a memory test on the gesture vocabulary. At the end of this section, some of the results, findings, and challenges of the designing of the gesture vocabulary are given. The second section of this chapter presents an evaluation on device recommendation system and scoring mechanism which is followed by a short discussion on the results and challenges of multimedia device recommendation system.

5.1 Gesture Vocabulary

First, we studied users' preferences and attitudes in the user survey. Then, we proposed a gesture vocabulary using the elicited information, which covers the essential commands of multimedia devices. Next, we evaluated the proposed vocabulary with the help of the user agreement test score similar to the one used by [66, 74]. Yet still remains the necessity of examining the vocabulary for our design principles: ease of performance and memorability. Moreover, a few points rose during the validation process. In this section, the evaluation procedure and results are given first then a brief discussion is presented.

5.1.1 Gesture Vocabulary Evaluation

We conducted a technical evaluation and an external user agreement test similar to [33, 47]. This evaluation included two parts: the first one consisting of an external user evaluation that also focused on technical evaluation and the second part consisting of the memorability evaluation of the proposed gestures. We used the combination of TV and Microsoft Kinect as an example of a gesture-enabled multimedia object to perform the technical evaluation.

Participants

Twenty participants volunteered for this study. Although all of them were university students or staff, they had various backgrounds such as engineering, science, management, social science, arts and linguistics. Two of them were left-handed. Participants were equally divided based on their gender (10 male and 10 female). They were asked to choose their age range between four options: teenage, twenties, thirties, forties or above. The average age range was between twenties and thirties (sd. = 6.07). Moreover, they originally came from diverse cultures and had as mother tongue various languages from all around the world. Also, only two of them either owned or frequently used a Microsoft Kinect sensor.

Apparatus

In order to check the feasibility of the proposed gesture vocabulary, a gesture recognition engine was developed. We used the Microsoft Kinect sensor due to its widespread availability and powerful Software Development Kit (SDK). The preliminary gesture recognition engine was implemented using the Dynamic Time Warping (DTW) technique [53] in which a developed engine normalizes the captured data before processing it. Upper body joint positions and angles in addition to hand area and dimension ratio are used as the searched-for features of gesture detection. The developed engine was embedded into a C# application that has three functionalities. First, it could playback the recorded videos for training the gestures. Second, participants could use it to practice the gestures. Third, participants used it to interact with predefined scenarios. The experiment room was equipped with the following items: a large screen TV (63in), a comfortable loveseat sofa, a Microsoft Kinect sensor and a Dell desktop computer (OPTIPLEX 760). Figure 5.1 shows the layout of this room.

Procedure

In the first part of this evaluation experiment, participants were asked to enter the room and sit on the sofa. First, the experimenter explained the participants' rights and experiment's privacy agreement. Next, the gesture-training step began. Participants watched four recorded videos: one for each group of referents (Table 3.1). Each video started with the name of the referent's group then the name of each command was shown and followed by the teaching video for said command. Gestures were taught both by audio and video. The experimenter answered and clarified any questions that were raised by the participant.

At the end of each video, participants were asked to face a screen where all of the presented gestures where listed on the right side and a live skeleton stream from the Kinect sensor played on the left side. Participants could try proposed gestures and imitate them using both movement feedback from the Kinect sensor and the visual indication for right gesture execution by the gesture recognition engine. Figure 5.2 (a) shows how each participant's skeleton was on the left and its corresponding referents were listed on the right. The skeleton image presented a real-time feedback to the participant in order

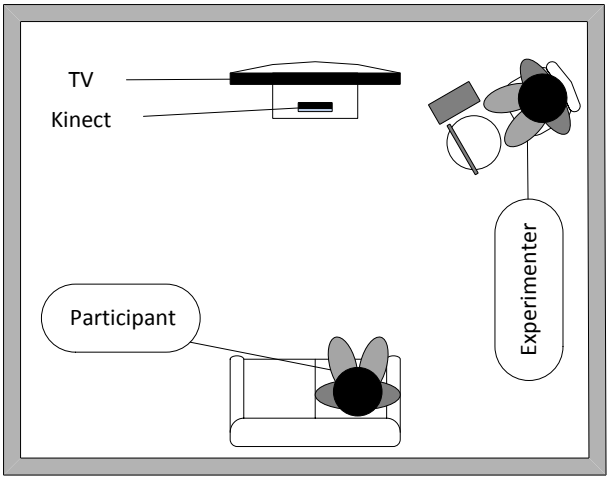
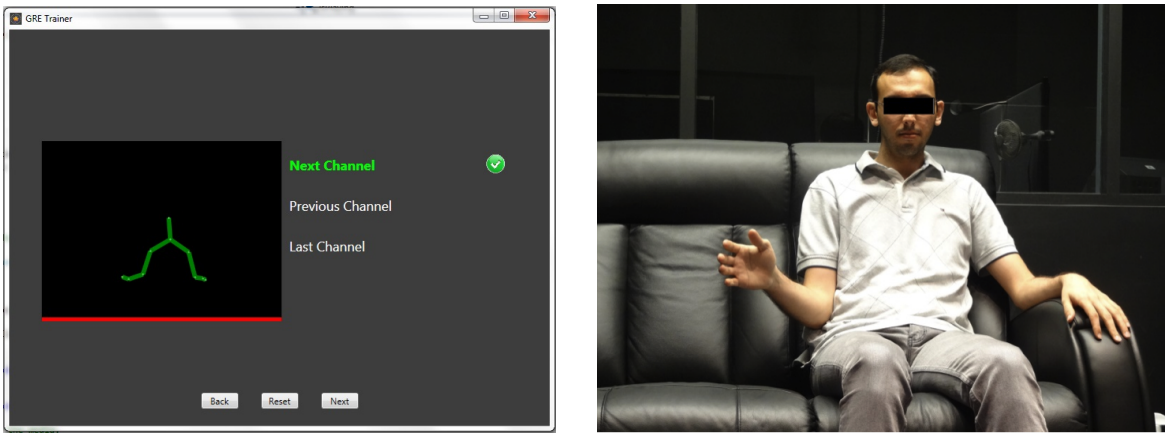


Figure 5.1: The layout the room for the lab-based experiment.

to help them clearly understand the captured movement. When the gesture recognition engine detected a gesture, it switched to a green bold format with a check mark in front of it. Figure 5.2 (b) shows a participant learning the relevant gestures. Participants could practice gestures a few times before informing the experimenter when they felt that they had mastered them, which usually took less than a minute.



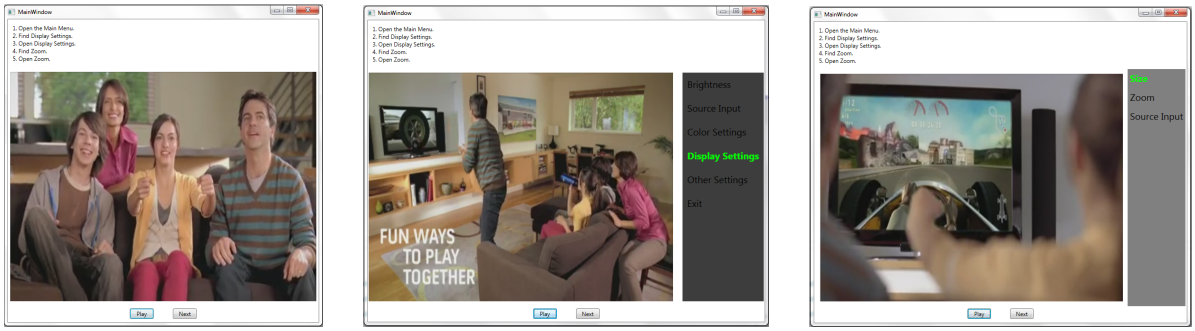
(a) Trainer application (b) Participant

Figure 5.2: The picture of a gesture training session.

After simulating of the gestures, they were asked to fill out a questionnaire about them. On the questionnaire they were asked to answer two 5-point Likert-scale questions. The first question asked if the gesture they had imitated was a good match for the command.

The next question asked them to rate the ease of performance of the gesture. Totally, they were taught 14 referents for 12 commands (Figure 3.1) where three commands had two referents. In those circumstances, participants had to answer another question about their preference on the two available options. This part took 15 to 20 minutes.

The next part of this experiment was designed to measure the memorability of proposed gestures and how intuitive they are. Participants were invited to return to the experiment room after two hours. Then, they were given four step-by-step scenarios to execute. They had to interact with a developed GUI and follow scenarios based on the response from the GUI. Figure 5.3 (a), (b), and (c) show three steps of the first scenario where participants were asked to interact with menu widgets. Also, Figure 5.3 (a) shows the main GUI layout, which is used in the second, third and fourth scenarios. The description of the scenarios is presented as follows.



(a) (b) (c)

Figure 5.3: The memorability test snapshots.

- **Scenario 1:** Participants were asked to open the main menu. Then, they were tasked with navigating through the menu to open the "Settings" submenu. Next, they had to find and open the "Zoom" option. In total, they had to perform 7 gestures, which covered all of the gestures on the "Menu Navigation" group.
- **Scenario 2:** Participants were asked to perform all "Channel Surfing" gestures with fixed time intervals. They had to follow this action order: Next Channel, Last Visited Channel and Previous Channel.
- **Scenario 3:** Participants were asked to perform "Volume Control" gestures in the third scenario. They followed this pattern: Increase Volume; Decrease Volume and Mute.
- **Scenario 4:** Participants were asked to turn off the TV.

In order to eliminate the effect of potential errors of the gesture recognition engine and reduce the complexity of this part of this experiment, the "Wizard of Oz technique was used. This is to say that participants were asked to imitate the gestures they had learned and explain them verbally. In doing so, they allowed the research results to focus only on

the memorability of the proposed gestures. The experimenter recorded the execution time and the number of mistakes for each scenario.

Post-Study Questionnaire Results

As explained in the previous subsection earlier, participants were asked to fill out a post-study questionnaire. Table 5.1 presents the results of this questionnaire. During the second phase of the user survey, "Next Channel" and "Previous Channel" had the highest scores. The results of this post-study questionnaire simply confirmed the results of the survey (section 3.1.2). Also, the results of the "Shutdown TV", "Up", and "Down" commands shows that they were well-accepted while "Snapping Fingers in order to open the main menu and the gestures for "Increase" or "Decrease" received a reasonable level of approval. Moreover, the "Hand on the Mouth" gesture achieved a very high score in this test. Two gestures were proposed for the "Open Submenu/Enter Option" command and both of them obtained a distribution within normal parameters. If the weakest options are eliminated in cases where two gestures exist for the same command, it can be found that "Last Visited Channel" has the lowest acceptance level among all of the studied commands. It can thus be inferred that it would be the least used of the commands.

Memorability Test Results

In the memorability test, three issues were examined. a) Discovering how quickly people can remember and use the gestures. This was achieved by measuring the execution time and gesture accuracy in the presented scenarios. Figure 5.4 shows the execution time of each scenario. As it can be seen, the first scenario has the widest distribution and it is has the longest duration. In this scenario, participants were expected to interact with the designed interface and performed at least 7 gestures. The average time for this scenario is 25.55 seconds while the median is 23.75 seconds. We judge this to be a reasonable time as participants were faced with the interface for the first time and also had to read and understand the instructions they were given before acting. The last scenario asked participants to shutdown the TV. As they only needed to perform a single intuitive gesture, i.e. a "Goodbye Wave, it was completed very quickly. The average time is 1.57 seconds while the median is only 1.5 seconds. This short response time confirms that people could remember the proposed gesture and perform it with great speed. Generally, the average time required to remember and perform a gesture is 3.13 seconds.

The next result of the memorability test was accuracy performance. The examiner recorded the total number of mistakes and the name of the gesture performed wrongly for each scenario (Table 5.2). The average of accuracy for all four scenarios equals to %91.54, which is satisfactory (especially the first one where participants had to perform a number of gestures according to the real-time responses of the system). Furthermore, since the median is 100% for all of scenarios, it can be concluded that more than half of our participants could perform the gestures required by each scenarios with complete accuracy. It is noted that the participants were not kept in the lab during the time interval between

Table 5.1: The results of the post-study questionnaire.

(a) Result of good match question

Command: Gesture	Good Match		
	Median	Mean	Std. dev.
Open Menu: Snapping Finger	4	4.2	0.77
Open Menu: Pushing Palm	3	3.25	1.12
Menu Navigation: Up / Down	5	4.55	0.51
Open Submenu / Enter Option: Pointing	4	3.95	0.76
Open Submenu / Enter Option: Push Pointing	4	3.65	0.99
Next / Prev. Channel: Swipe Hand Left/ Right	5	4.6	0.50
Last Visited Channel: Circular Movement of Hand	4	3.65	1.04
Increase / Decrease Volume: Moving Up / Down	4.5	4.35	0.81
Mute: Hand on Mouth	5	4.3	1.03
Mute: Closing Palm	4	3.85	1.04
Shutdown TV: Goodbye Wave	5	4.55	0.51

(b) Result of ease of performance question

Command: Gesture	Ease of Performance		
	Median	Mean	Std. dev.
Open Menu: Snapping Finger	4	4.35	0.59
Open Menu: Pushing Palm	4	4	0.86
Menu Navigation: Up / Down	4	4.25	0.79
Open Submenu / Enter Option: Pointing	4	4.2	0.77
Open Submenu / Enter Option: Push Pointing	4	3.8	0.89
Next / Prev. Channel: Swipe Hand Left/ Right	4	4.35	0.59
Last Visited Channel: Circular Movement of Hand	4	3.7	1.03
Increase / Decrease Volume: Moving Up / Down	5	4.5	0.61
Mute: Hand on Mouth	5	4.4	0.82
Mute: Closing Palm	4	4.2	0.83
Shutdown TV: Goodbye Wave	5	4.55	0.60

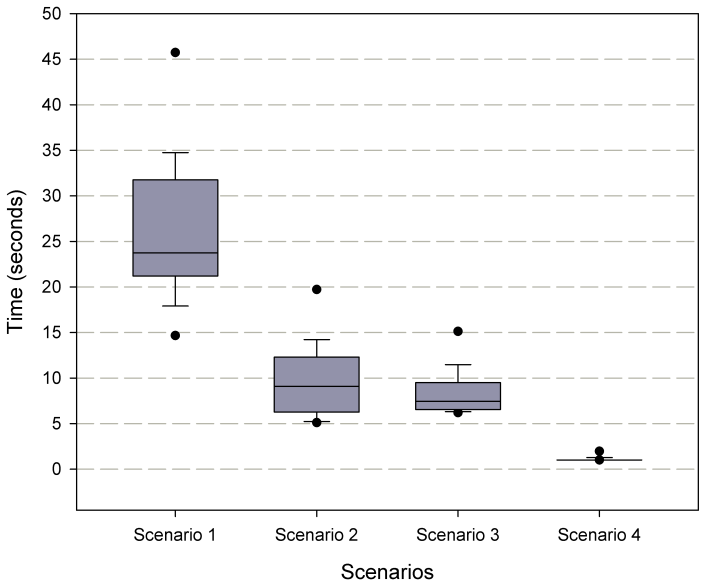


Figure 5.4: Statistics of the execution time in each scenario.

training and testing. They left the room after finishing the training part of the experiment and came back after two hours.

Table 5.2: The accuracy results of the memorability test.

Scenario	Median	Average	Std. dev.
Scenario 1	100%	92.48%	0.15
Scenario 2	100%	87.72%	0.23
Scenario 3	100%	96.49%	0.15
Scenario 4	100%	89.47%	0.31

c) The third result was concerned with eliciting the connection between participants' preferences and memories (Table 5.3). In order to do this test, only three commands ("Open Menu", "Open Submenu/Enter Option" and "Mute") were considered from all those that possessed two options. These commands were selected due to the closeness of their scores according to the second survey's results). In this questionnaire the participants were asked to choose one gesture as their preferred gesture for each associated command. Table 5.3 presents the results of those three questions. It must be noted that it was also recorded which gesture participants performed during the scenarios. To open the main menu, "Snapping Fingers was reported as being preferred three times more than "Pushing Palm. However, on the scenarios, the result was different as "Pushing Palm was used slightly more than "Snapping Finger. The reason behind this huge difference could associate with users' experience of opening the menus of other programs by pushing the menu button. It can cancel out being both a better match and an easier gesture to recall

when participants needed to make real-time decisions. However, the user preferences on the two other commands were confirmed by the memorability test since they had almost the same results. Although both of proposed gestures for these gestures did not have a very strong advantage over each other, it indicates that the embodied cognition experience through imitating those gestures might help participants better remember their preferred gesture after a time.

Table 5.3: The comparison between the available options of three commands.

Command	Gesture	No. of Preferred	No. of Executed
Open Menu	Snapping Finger	15	9
	Pushing Palm	5	11
Open Submenu/Enter Option	Pointing	13	11
	Push Pointing	7	9
Mute	Hand on Mouth	13	13
	Closing Palm	7	7

5.1.2 Gesture Vocabulary Discussion

This section presents a brief discussion about the results achieved and the challenges confronted during the study. First, the preliminary study is discussed. Next, certain points of note in the online questionnaire are reviewed. Finally, a discussion on the technical evaluation is given.

Preliminary Study

There are two significant issues related to the first phase of the user study. The first issue dealt with the size of and the motivations behind the nominated commands. Q21 in Table 3.4 confirmed the results of [66] that indicated that people might still prefer to use handheld controllers when they are available. As a result, we decided to choose a small, frequently used set of commands. It is noted that while twelve referents were initially selected for this study, two shared the same gesture and were thusly merged together. By taking a close look at the four groups of commands, the need for specific commands such as "Override" and "Volume Control" became more evident. The "Menu Navigation" group contains only the fundamental and necessary commands for menu interaction such as opening, navigating, and entering. The "Channel Surfing" group, comprised of three commands, holds two of the most essential functions, "Next Channel" and "Previous Channel", both of which change the channel by one unit per gesture. There is no referent that changes the channel for more than 1 unit per gesture but users can open the menu and use search module or virtual keypad to change the channel for more than 1 unit. The "Last Visited" channel was added to the vocabulary to help users with turning back to their last visited channel without using the menu or widgets.

The second issue in this phase was raised by notes of participants. When they were asked to suggest gestures for "Next/Previous Channel", they usually suggested horizontal

movements. They all provided a similar example, which was an action likening to that of turning a page in a book. While this was a good suggestion, which accurately shows the unconscious ideas and mental models of a communal mind, it raised the issue of culture and language. A person from a specific country uses different directional indicators, which can be opposite to those from a different country (example: Canada turns pages left to right but Japan does so right to left). We determined that the direction of movement should be adjustable by the users. This challenge was made evident in the last steps of this study when we asked participants to change the channels. A few of them used the gesture right to left instead of left to right, largely due to the fact that their native language is written right to left. Moreover, gestures such as putting a hand over the mouth, which was proposed for the "Mute" command, could potentially have different meanings in different cultures. This issue is a well-known problem in HCI and has been addressed in similar studies [33, 47, 51, 55, 66, 74].

Online Questionnaire

Participants of the preliminary study were given open-ended questions and, as a result, their-suggestions involved many parts of the body, including hands, head, chest, eyes, ears, and mouth. As shown in Table 3.4, gestures that utilize both hands (e.g. clapping) were not selected. From this, we can infer that one-handed gestures are generally preferred over two-handed gestures. This also confirms similar results from [66, 74].

In addition, participants of the preliminary study had a background offering less technical IT knowledge than those in the second group. "Turning the Head to Left/Right" and "Focusing in one Direction" are some suggested gestures in the preliminary study. They are especially good examples to show that participants were not focused on technical aspects. Instead, they gave their suggestions based on instinct and intuition. The second group of participants used their knowledge consciencelessly when selecting among the options given. As a result, gestures that might be difficult to detect from a long distance using depth technology (e.g. focusing) or hard to understand (e.g. turning head) were not chosen. It was interesting to note that a majority of participants (67%) identified that they chose their answers because they felt more intuitive. Results of the external validation and memorability tests verified this answer even though the participants were completely different.

Technical Evaluation and User Agreement

Results of the online questionnaire do not show a unique answer for "Open Menu", "Open Submenu/Enter Option" and "Mute". To determine the final gesture for those three referents, the top two answers for each command were selected and taught to the technical evaluation test's participants. Participants were then asked to choose one of them. At the end of this part of the presented study, each of the three referents had a distinctive answer. Training and an imitating process of technical evaluation might assist participants in clarifying which gesture is more intuitive without being biased by their ease of performance. For instance, the "Mute" command could be a good example. Initially, in the

online questionnaire, the "Closing Fist" gesture was slightly ahead of "Putting Hand over the Mouth". However, at the end of the last step, "Putting Hand over the Mouth" had a higher success rate both on the post-study questionnaire and the memorability test. This indicates that although "Closing Fist" is easier to perform and might theoretically be a better answer for the mute command, after training and mimicking, "Putting Hand over the Mouth" seemed more intuitive to participants.

We should also highlight that describing a gesture without any visual element and imitation may lead the gesture to be a concept rather than an interaction model. When participants were asked to learn and perform these gestures, embodied cognition became involved in the decision and memorization process [72]. We believe that this could be used to explain the differences in the results of the online questionnaire and technical evaluation for the three commands with two options. Moreover, embodied cognition might be the prominent element for the high approval rates of all the referents and exceptionally great success rates on the defined scenarios [58]. Finally, although we examined intermediate term memory with duration from minutes to hours [27] in our memorability test, most of the participants indicated that they could remember the proposed gestures even a week after having learned them. They could also perform the gestures correctly. This could be considered as a positive feedback for both the right size of the vocabulary set and the intuitiveness of the proposed referents.

Moreover, participants in the technical evaluation had two more concerns. First, they said that they wanted to be able to perform a continuous interaction with devices. For example, they wanted to change a few channels at a time by using one swipe gesture just like scrolling. Generally, this concern was mentioned for gestures that consist of a horizontal or vertical hand motion. We believe that, though this concern is important, it is more related to the implementation of the gesture vocabulary than to the characteristics of it. The second concern of participants was about the engagement mechanism. It is a controversial subject in gesture and speech-enabled systems. Defining a simple and easy engagement mechanism may not be accurately detectable by the current technology (e.g. direction of eyes). On the other hand, defining a mechanism that can be precisely detected by the available sensor technology may not be easy to perform for the users. However, advances in technology will help researchers find a solution for this problem that will be both realistic for our current technological levels and user friendly. For example, the next generation of Microsoft Kinect will be more advanced and will be able to detect movement of fingers more accurately than the present one. This could help us reduce the margin of error of our presently developed gesture recognition engine and design an acceptable engagement mechanism.

5.2 Media Redirection

5.2.1 Media Redirection Evaluation

To evaluate the device recommendation system, we defined 4 scenarios and conducted a user survey. In these scenarios, participants were given access to all types of devices in

Table 5.4: The results of the evaluation study.

Method		scenario 1	scenario 2	scenario 3	scenario 4	Avg.
Our Method	1^{st} suggestion	90%	30%	70%	90%	70%
	$1^{st}/2^{nd}$ suggestion	100%	80%	80%	100%	90%
Distance Only	1^{st} suggestion	90%	30%	60%	90%	67.5%
	$1^{st}/2^{nd}$ suggestion	100%	70%	70%	100%	85%
AirPlayer [61]	–	100%	30%	30%	90%	62.5%

their effective distance. The content's length was the only variable in our scenarios. In the first scenario, it was 2 min while it was 7 min, 22 min, and 90 min in the second, third and fourth scenarios respectively. The user had to select his/her preferred device for watching a 7 min video while he/she has access to a smartphone in 0.25 m, tablet in 0.75 m, personal computer in 2.5 m, and TV in 5 m. In total, 10 people participated in our study with their ages ranging from 20 to 40. The users all had different professions and media playback habits. We compared the user responses to the device recommendations made by the proposed user interface to measure the usability.

We also compared the proposed approach with two other methods. The first one only used the distance between the device and the user to suggest a new device. As such, the method always suggests the device that whose most effective area is closest to the user but also takes into account the same learning mechanism as ours. The second method is AirPlayer [61]. AirPlayer only supports intimate and public devices. Furthermore, it does not possess the learning mechanism we used for other devices and so cannot adapt itself to users' preferences. Table 5.4 shows the results of this evaluation study.

In this study, the total accuracy of the first suggestion for our system is equal to 70% while it was 67.5% for the distance only method and 62.5% for the AirPlayer. But, since the proposed method can learn from the users' responses and adapt itself to their preferences, we decided to study the accuracy of these methods after two interactions. This is not applicable to the AirPlayer due to its lack of learning mechanism. The total accuracy after two interactions for our system was 90%, however it was 85% for the distance only method. To summarize, we can say that our proposed method had a higher accuracy than other similar methods in these scenarios. As a result, it could provide an environment that has more successful proxemic multimedia interactions.

5.2.2 Media Redirection Discussion

There are three points that should be discussed regarding the proposed system. First, the proposed solution checks the identity of users by tracking their intimate devices. Then, it uses their personal information to recommend a device to them and adaptively updates itself for each user. Also, each device or user updates the location information when it has a movement. The devices are recommended based on the distance between devices. Therefore, we can conclude that the proposed solution directly uses 3 out of 5 of the proxemic interaction dimensions that were introduced earlier by Greenberg et al [18]. The

remaining two dimensions can be involved in this system as well. The orientation of the user and location of devices may influence the users' decision on which playback device to use.

Second, we should compare the proposed UI with the definition of proxemic interaction by Hall [21]. The proposed UI is designed to work over the IoT. As such, it can handle proxemic interactions in multiple rooms (i.e. macro-level). Yet, it provides an environment for micro-level proxemic interactions as well. Devices within a single room can interact with each other when a measured change occurs in any of the 3 mentioned proxemic dimensions.

The last point concerns the engagement mechanism. Since the number of online devices is growing very fast, most of the times users are surrounded by a large number of devices. However, since it is not easy to migrate from one device to the other, users may not change the device that they are using even if it is contextually appropriate. However, the proposed UI considers the users' preferences and proxemic information to suggest a new device, which can increase the number of accepted recommendations. Also, it can assist users through the migration process by handling redirection process in the background. Hence, the proposed UI facilitates the engagement process for the users and provides them more with options to use.

Chapter 6

Conclusion and Future Works

This chapter starts a brief conclusion of this study based on three perspectives: user engagement mechanisms, device recommendation systems, and gesture vocabulary. Then, it presents a short discussion about possible future work in ProGes.

6.1 Conclusion

We proposed ProGes, a Proxemic and Gesture interaction system for multimedia devices over the IoT. ProGes has three major advantages that make it a new technology with a lot of potential. First, it is a task-centered system that targets the user interaction with multimedia devices over the IoT. Second, it is designed based on broad user surveys and elicitation studies, which results in ProGes being a user-centered system. Third, although it delivers a distributed user interface, it has a cloud-based control unit that provides a reliable support to different parts by coordinating the interactions. To design the ProGes, we designed and implemented different algorithms for a few existing problems in the domain of user interface design.

6.1.1 User Engagement Mechanism

Multimedia devices have independent user interfaces and work separately. Therefore, users usually have to engage with such devices explicitly. However, ProGes suggests a new engagement mechanism that encourages the user to make use of new devices. In this system, proximity information is used to initiate users to new devices. This is an implicit method that always tracks users so that whenever they enter into the perimeter of a new device, they recommend the said device. Users can then accept this recommendation and take the control of the new device. A good application of this mechanism is for gesture-enabled devices where it is hard to discern whether a person is user or not. Using this mechanism, ProGes comes to know the proximity information of users and, in doing so, can detect them amongst other people in order to provide them with an exemplary gesture control service.

6.1.2 Device Recommendation System

ProGes is designed for user and multimedia devices interactions over the IoT. In the IoT, users are usually surrounded by multiple devices. Hence, ProGes proposes a device recommendation system that facilitates the interaction between users and devices. We conducted a user study to elicit users' preferences towards using multimedia devices and also reviewed the literature to extract the influential elements on users' device preferences. We found three groups of factors: multimedia content properties, users' personal information and device capabilities and proxemic information. Using the combination of these elements, the proposed system gives a score between 0 and 1 to each device. Then, the device that has the highest score is recommended to the user and the necessary media redirection is handled by ProGes should the user accept the recommendation. The proposed algorithm adaptively trains itself towards users' attitude over time based on the users' feedback. The scoring mechanism is validated in four scenarios by a user study and it has the acceptable average accuracy of 70%. Using the feedback from the acceptance ratio, the system trains itself to anticipate users' attitudes over time and has the potential to reach an average accuracy of 90%.

6.1.3 Gesture Vocabulary

Since the arrival of recent advances in gesture recognition systems specifically those relating to depth-based gesture recognition researchers have suggested different general and task-oriented gesture vocabularies. However, these vocabularies have been failed to become universal and support different devices due to their limitations on platform requirements. ProGes introduces a gesture vocabulary for a set of multimedia device control commands. This vocabulary was designed to be simple, easy to remember and intuitive. We conducted a broad two-step user study to be able to propose said vocabulary. During the first step, users' preferences and attitudes were elicited and used to design the vocabulary. The average of the agreement score of the proposed vocabulary in this step of the user survey exceeded the average score of similar studies. The second step validated the proposed vocabulary by an external user evaluation study. Said test attained high scores on the post-study questionnaire and offered exceedingly accurate performances on the memorability test. Although the size of the proposed vocabulary is smaller than in similar studies, it has the same functionality. These achievements can prove that ProGes has accomplished its goals as well as reaching a better agreement score than many of its competitors.

6.2 Future Works

There exist a few avenues of possible improvement for ProGes that could be addressed in future works. The first one is the implementation of ProGes in such a way that it will include more devices and interaction modalities, which will build a generic, distributed user interface for IoT. This is a huge step and it could move the proposed system one step forward towards being used in everyday life. Also, it could reveal some possible

hidden problems that would need to be addressed, which is very beneficial to us in that it encourages innovation and progress.

The other possible future work lies in developing a more accurate and adjustable gesture recognition engine for the proposed gesture vocabulary. This can be done using the new generation of Microsoft Kinect, which has a very high level of accuracy and resolution. Using such an engine, it would be possible to study the effect of the confronted challenges (e.g. cultural and lingual) in more details. In addition to that, the proposed vocabulary could be evaluated from other perspectives such as physiological, etc.

Last, finding and involving more elements in the scoring mechanism may improve the device recommendation system. Also, the effects of the personalization process in the device recommendation system could be studied more precisely by conducting a long-term user evaluation study.

APPENDICES

Appendix A

Statistical Analysis of Gesture Elicitation Study

We applied a statistical analysis on the samples to describe the overall populations, regarding preferences and attitudes. The questions in the survey are divided into three groups: Q(1-2) as the basic grouping criterion (independent variable); Q(3-4, 17-22) as the gesture control attitude (dependent variable); Q(5-16) as the gesture preference or suggestion (dependent variable). Specifically, we used the independent-samples *T-test*, and the bivariate correlation to compare the gesture attitude between gender/age and gesture preference.

The independent-samples *T-test* was extended for equality of means to measure the significance of the difference on gesture control attitude between the male and female group. Table A.1a[1] produces two tests of the difference between the two groups. The first test assumes that the variances of the two groups are equal. The *Levene's test* is used for equality of variance to measure this assumption. Here, the significance value in Q(4, 17, 18, 19) is greater than 0.10, thus it can be assumed that the two groups in these questions have equal variances. Based on the results of the equality test of variances, the *T-test* is applied to find out if the two groups (male/female) have differences in attitude on gesture control. It was found that in Q(4), the significance value of the test is less than 0.05, indicating that for Question 3 and 4, there is no difference between the male and female groups. I.e., the male play video games (especially that with gesture control) as frequently as the female. Moreover, the male group has more positive answer in the Question (17, 18, 19) indicating the male is more likely to try the gesture control.

Similarly, the attitude difference was compared between the age groups. We divided the samples were divided into two groups according to age: $\leq$25 years old (young group) and >25 years old (old group). The significance values of Q(17, 18, 19, 20, 21) are greater than

[1] *indicates the significance value larger than 0.10 in Levene statistics. **indicates the significance value less than 0.05 in t test for equality of means. The t column displays the t statistics for each question, this is calculated as the ratio of the difference between sample means divided by the standard error of the difference. the *Sig.* column displays a probability from the t distribution, which is the probability of obtaining an absolute value greater than or equal to the observed t statistics, if the difference between the sample means is purely random.

0.10 and thus pass *Levene's* test for Equality of variances. Hence, it was finally obtained that there is no attitude difference in Q(21) on gesture control between the two age groups indicating both the age groups are optimistic in trying the gesture control compared with the conventional remote control. The young group has more positive answer in the Question (17, 19, 20) indicating that they prefer to remember more gesture control commands to increase control accuracy. Whereas, the old group is more optimistic in Question (18) which means the old people prefer the gesture vocabulary control idea more than the young group. Table A.1b[2] shows the complete results of these tests.

Table A.1: Results of T-test on Gender and Age Difference

(a) T-test on Gender

Question	Levene's Test for Equality of Variances		T-test for Equality of Means		Mean	
	F	Sig.	t	Sig.	Male Group	Female Group
Q3	6.771	0.011	-4.531	**0.000 ****	3.0682	4.1892
Q4	3.395	**0.69 ***	-2.539	**0.013 ****	3.9773	4.5405
Q17	0.117	**0.734 ***	-0.444	0.658	2.2273	2.3243
Q18	1.128	**0.721 ***	-0.134	0.894	2.6136	2.6486
Q19	0.872	**0.353 ***	-0.803	0.425	2.1818	2.3784
Q20	5.725	0.019	1.359	0.178	1.4773	1.2703
Q21	0.096	0.758	-0.776	0.440	2.2727	2.4595
Q22	0.940	0.335	0.764	0.447	1.8636	1.6486

(b) T-test on Age Difference

Question	Levene's Test for Equality of Variances		T-test for Equality of Means		Mean	
	F	Sig.	t	Sig.	Young Group	Old Group
Q3	7.465	0.008	1.698	0.096	3.7600	3.2333
Q4	21.020	0.000	2.253	**0.031 ****	4.4600	3.8333
Q17	1.472	**0.229 ***	-0.264	0.792	2.2400	2.3000
Q18	0.012	**0.915 ***	0.811	0.420	2.7200	2.5000
Q19	0.355	**0.553 ***	-0.499	0.619	2.2400	2.3667
Q20	0.402	**0.528 ***	-0.450	0.654	1.3600	1.4333
Q21	0.767	**0.384 ***	-2.189	**0.032 ****	2.1400	2.6667
Q22	10.044	0.002	-1.823	0.075	1.5600	2.1333

This appendix also shows the correlation between difference gesture preference Q(5-16). The Bivariate correlation was extended to compute the Pearson's correlation coefficient, it is shown how the variables or rank orders are related, and display the results in a matrix, shown in Table A.2[3]. Negative values of the Pearson correlation are present in the test, this happens due to some outliers in the distribution. However, in the present case, as the

[2] 1

[3] *indicates the correlation is significant at the 0.05 level (2-tailed).**indicates the correlation is significant at the 0.01 level (2-tailed).

significance values corresponding with the negative correlations are greater than 0.10, the correlation is not sufficiently trustworthy, and hence can be ignored. Among the trustworthy correlation relation whose significance level is at 0.01/0.05 level (emphasized as */**), the following question pairs have a high correlation level (more than 0.7): Q5↔Q6, Q8↔Q9, and Q12↔Q13. This means that the gesture preferences for the following referents always happened in pairs: next channel↔previous channel, more volume↔less volume, menu (up in the menu)↔menu (down in the menu), which is not surprising.

Table A.2: Correlation Test of Gesture Preferences

		Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Q5	Correlation	1	***0.732 *****	-0.011	0.142	0.128	0.153	0.083	0.017	0.113	0.083	-0.014	-0.039
	Sig.		0.000	0.921	0.206	0.256	0.172	0.461	0.881	0.316	0.461	0.902	0.728
Q6	Correlation	***0.732 *****	1	0.132	0.103	0.082	0.197	0.183	0.067	0.132	0.052	0.053	0.060
	Sig.	0.000		0.239	0.359	0.467	0.078	0.102	0.552	0.239	0.646	0.638	0.592
Q7	Correlation	-0.011	0.132	1	-0.046	-0.117	-0.007	0.064	0.215	0.148	0.054	0.133	0.116
	Sig.	0.921	0.239		0.684	0.300	0.949	0.572	0.054	0.189	0.629	0.236	0.301
Q8	Correlation	0.142	0.103	-0.046	1	***0.891 *****	0.053	0.194	0.156	***0.253 ****	-0.031	0.147	0.173
	Sig.	0.206	0.359	0.684		0.000	0.642	0.083	0.165	0.022	0.784	0.190	0.122
Q9	Correlation	0.128	0.082	-0.117	***0.891 *****	1	0.076	***0.275 ****	0.111	***0.228 ****	-0.064	***0.235 ****	0.114
	Sig.	0.256	0.467	0.300	0.000		0.502	0.013	0.324	0.041	0.572	0.035	0.312
Q10	Correlation	0.153	0.197	-0.007	0.053	0.076	1	***0.226 ****	***0.223 ****	0.194	0.174	0.119	***0.228 ****
	Sig.	0.172	0.078	0.949	0.642	0.502		0.043	0.045	0.083	0.121	0.291	0.041
Q11	Correlation	0.083	0.183	0.064	0.194	***0.275 ****	***0.226 ****	1	***0.350 *****	***0.350 *****	0.011	***0.259 ****	0.197
	Sig.	0.461	0.102	0.572	0.083	0.013	0.043		0.001	0.001	0.924	0.019	0.078
Q12	Correlation	0.017	0.067	0.215	0.156	0.111	***0.223 ****	***0.350 *****	1	***0.917 *****	-0.105	***0.278 ****	0.100
	Sig.	0.881	0.552	0.054	0.165	0.324	0.045	0.001		0.000	0.353	0.012	0.375
Q13	Correlation	0.113	0.132	0.148	***0.253 ****	***0.228 ****	0.194	***0.350 *****	***0.917 *****	1	-0.120	***0.322 *****	0.100
	Sig.	0.316	0.239	0.189	0.022	0.041	0.083	0.001	0.000		0.287	0.003	0.376
Q14	Correlation	0.083	0.052	0.054	-0.031	-0.064	0.174	0.011	-0.105	-0.120	1	0.099	-0.004
	Sig.	0.461	0.646	0.629	0.784	0.572	0.121	0.924	0.353	0.287		0.378	0.974
Q15	Correlation	-0.014	0.053	0.133	0.147	***0.235 ****	0.119	***0.259 ****	***0.278 ****	***0.322 *****	0.099	1	0.174
	Sig.	0.902	0.638	0.236	0.190	0.035	0.291	0.019	0.012	0.003	0.378		0.121
Q16	Correlation	-0.039	0.060	0.116	0.173	0.114	***0.228 ****	0.197	0.100	0.100	-0.004	0.174	1
	Sig.	0.728	0.592	0.301	0.122	0.312	0.041	0.078	0.375	0.376	0.974	0.121	

References

[1] Ekahau wifi tags and badges. http://www.ekahau.com/real-time-location-system/technology/wi-fi-tags, 2012.

[2] C. Ackad, A. Clayphan, R.M. Maldonado, and J. Kay. Seamless and continuous user identification for interactive tabletops using personal device handshaking and body tracking. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 1775–1780, New York, NY, USA, 2012. ACM.

[3] M. Annett, T. Grossman, D. Wigdor, and G. Fitzmaurice. Medusa: A proximity-aware multi-touch tabletop. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 337–346, New York, NY, USA, 2011. ACM.

[4] P. Bahl and V.N. Padmanabhan. RADAR: an in-building rf-based user location and tracking system. In *INFOCOM 2000, 19th Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 775–784, 2000.

[5] X. Bai and L.J. Latecki. Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1282–1292, 2008.

[6] B. Bauer and H. Hienz. Relevant features for video-based continuous sign language recognition. In *Proceedings of the 4th International Conference onAutomatic Face and Gesture Recognition*, pages 440–445, 2000.

[7] K. Bing, L. Fu, Y. Zhuo, and L. Yanlei. Design of an internet of things-based smart home system. In *Intelligent Control and Information Processing (ICICIP), 2011 2nd International Conference on*, volume 2, pages 921–924, Jul 2011.

[8] P. Breuer, C. Eckes, and S. Muller. Hand gesture recognition with a novel IR time-of-flight range camera: a pilot study. *Proceedings of the 3rd International Conference on Computer Vision/Computer Graphics Collaboration Techniques*, pages 247–260, 2007.

[9] G. Broll, E. Rukzio, M. Paolucci, M. Wagner, A. Schmidt, and H. Hussmann. Perci: Pervasive service interaction with the internet of things. *Internet Computing, IEEE*, 13(6):74–81, Nov 2009.

[10] G. Broll, S. Siorpaes, E. Rukzio, M. Paolucci, J. Hamard, M. Wagner, and A. Schmidt. Supporting mobile service usage through physical mobile interaction. In *Pervasive*

Computing and Communications, 2007. PerCom '07. Fifth Annual IEEE International Conference on, pages 262–271, Mar 2007.

[11] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S.A. Shafer. Easyliving: Technologies for intelligent environments. In *Proceedings of the 2nd International Symposium on Handheld and Ubiquitous Computing*, HUC '00, pages 12–29, London, UK, UK, 2000. Springer-Verlag.

[12] S. Celebi, A.S. Aydin, T.T. Temiz, and T. Arici. Gesture recognition using skeleton data with weighted dynamic time warping. *Computer Vision Theory and Applications. Visapp*, 2013.

[13] S. Connell, P.Y. Kuo, L. Liu, and A.M. Piper. A wizard-of-oz elicitation study examining child-defined gestures with a whole-body interface. In *Proceedings of the 12th International Conference on Interaction Design and Children*, pages 277–280, 2013.

[14] A. Corradini and H.M. Gross. Camera-based gesture recognition for robot control. In *Proceedings of IEEE International Conference on Neural Networks*, pages 133–138, 2000.

[15] J. Crowley, F. Berard, and J. Coutaz. Finger tracking as an input device for augmented reality. In *Proceedings of the International Workshop on Gesture and Face Recognition*, pages 195–200, 1995.

[16] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133, 2000.

[17] C. Drane, M. Macnaughtan, and C. Scott. Positioning GSM telephones. *Communications Magazine, IEEE*, 36(4):46–54, 59, Apr 1998.

[18] S. Greenberg, N. Marquardt, T. Ballendat, R. Diaz-Marino, and M. Wang. Proxemic interactions: The new ubicomp? *Interactions*, 18(1):42–50, Jan 2011.

[19] Y. Gu, A. Lo, and I. Niemegeers. A survey of indoor positioning systems for wireless personal networks. *IEEE Communications Surveys Tutorials*, 11(1):13–32, Jan 2009.

[20] D. Guinard and V. Trifa. Towards the web of things: Web mashups for embedded devices. In *In MEM 2009 in Proceedings of WWW 2009*. ACM, 2009.

[21] E.T. Hall. *The hidden dimension*. Anchor Books New York, 1969.

[22] D. Hart. A brief history of NSF and the Internet. http://www.nsf.gov/od/lpa/news/03/fsnsf_internet.htm, Aug 2003.

[23] J. Hightower and G. Borriello. Location sensing techniques. Technical report, 2001.

[24] J. Hightower and G. Borriello. A survey and taxonomy of location systems for ubiquitous computing. *IEEE computer*, 34(8):57–66, Aug 2001.

[25] D.D. Hoffman and B.E. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 42(3):195–204, 1982.

[26] W. Ju, B.A. Lee, and S.R. Klemmer. Range: Exploring implicit interaction through electronic whiteboard design. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 17–26, New York, NY, USA, 2008. ACM.

[27] R.P. Kesner. Neurobiological views of memory. In Raymond P. Kesner and Joe L. Martinez, editors, *Neurobiology of Learning and Memory*, pages 271–304. Academic Press, Burlington, second edition edition, 2007.

[28] S. Kim, C.B. Park, and S.W. Lee. Tracking 3d human body using particle filter in moving monocular camera. In *18th International Conference on Pattern Recognition*, volume 4, pages 805–808, 2006.

[29] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg. Compass: A probabilistic indoor positioning system based on 802.11 and digital compasses. In *Proceedings of the 1st international workshop on Wireless network testbeds, experimental evaluation and characterization*, pages 34–40. ACM, 2006.

[30] G. Kortuem, F. Kawsar, D. Fitton, and V. Sundramoorthy. Smart objects as building blocks for the internet of things. *Internet Computing, IEEE*, 14(1):44–51, Jan 2010.

[31] M. Kranz, P. Holleis, and A. Schmidt. Embedded interaction: Interacting with the internet of things. *Internet Computing, IEEE*, 14(2):46–53, Mar 2010.

[32] J. Krumm and K. Hinckley. The NearMe wireless proximity server. In *UbiComp 2004: Ubiquitous Computing*, pages 283–300. Springer, 2004.

[33] C. Kühnel, T. Westermann, F. Hemmert, S. Kratz, A. Müller, and S. Möller. I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies*, 69(11):693–704, 2011.

[34] M. Laibowitz, J. Gips, R. AyIward, A. Pentland, and J.A. Paradiso. A sensor network for social dynamics. In *Information Processing in Sensor Networks. IPSN 2006. The 5th International Conference on*, pages 483–491, 2006.

[35] H.K. Lee and J.H. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):961–973, 1999.

[36] J. Lee and T.L. Kunii. Model-based analysis of hand posture. *IEEE Computer Graphics and Applications*, 15(5):77–86, 1995.

[37] Z. Li, J. Lin, M.I. Akodjenou, G. Xie, M.A. Kaafar, Y. Jin, and G. Peng. Watching videos from everywhere: A study of the pptv mobile vod system. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, pages 185–198. ACM, 2012.

[38] H. Liu, H. Darabi, P. Banerjee, and J. Liu. Survey of wireless indoor positioning techniques and systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(6):1067–1080, Nov 2007.

[39] K. Luyten and K. Coninx. Distributed user interface elements to support smart interaction spaces. pages 277–286. IEEE, 2005.

[40] N. Marquardt, R. Diaz-Marino, S. Boring, and S. Greenberg. The proximity toolkit: Prototyping proxemic interactions in ubiquitous computing ecologies. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 315–326, New York, NY, USA, 2011. ACM.

[41] N. Marquardt, K. Hinckley, and S. Greenberg. Cross-device interaction via micro-mobility and f-formations. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 13–22, New York, NY, USA, 2012. ACM.

[42] F. Mattern and C. Floerkemeier. From the Internet of Computers to the Internet of Things. In *From active data management to event-based systems and more*, pages 242–259. Springer, 2010.

[43] M.R. Morris. Web on the wall: insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, pages 95–104, 2012.

[44] M.R. Morris, J.O. Wobbrock, and A.D. Wilson. Understanding users' preferences for surface gestures. In *Proceedings of Graphics Interface 2010*, pages 261–268, 2010.

[45] M.A. Nacenta, Y. Kamber, Y. Qiang, and P.O. Kristensson. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1099–1108, 2013.

[46] Cisco News. Connections counter: The Internet of everything in motion. http://newsroom.cisco.com/feature-content?type=webcontent&articleId=1208342, Jul 2013.

[47] M. Nielsen, M. Stoerring, T.B. Moeslund, and E. Granum. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *Proceedings of the Gesture Workshop*, pages 409–420, 2003.

[48] H.S. Park, E.Y. Kim, S.S. Jang, S.H. Park, M.H. Park, and H.J. Kim. *Pattern Recognition and Image Analysis*, chapter HMM-Based Gesture Recognition for Robot Control, pages 607–614. Springer, 2005.

[49] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *Proceedings of the CHI '13 Extended Abstracts on Human Factors in Computing Systems*, pages 955–960, 2013.

[50] I. Ramani, R. Bharadwaja, and P.V. Rangan. Location tracking for media appliances in wireless home networks. In *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, volume 2, pages 769–772, Jul 2003.

[51] M. Rehm, N. Bee, and E. André. Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, volume 1, pages 13–22. British Computer Society, 2008.

[52] Z. Ren, J. Yuan, and Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1093–1096, 2011.

[53] M. Reyes, G. Dominguez, and S. Escalera. Featureweighting in dynamic timewarping for gesture recognition in depth data. In *Proceedings of the Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1182–1188, 2011.

[54] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, 1994.

[55] J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 197–206, 2011.

[56] B. Sanou. *ICT Facts and Figures*. International Telecommunications Union, Geneva, Switzerland, Feb 2013.

[57] D. Schuler and A. Namioka. *Participatory Design: Principles and Practices*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1993.

[58] C.L. Scott, R.J. Harris, and A.R. Rothe. Embodied cognition through improvisation improves memory for a dramatic monologue. *Discourse Processes*, 31(3):293–305, 2001.

[59] T. Seyed, C. Burns, M. Costa Sousa, F. Maurer, and A. Tang. Eliciting usable gestures for multi-display environments. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, pages 41–50, 2012.

[60] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[61] H. Sørensen, M.G. Kristensen, J. Kjeldskov, and M.B. Skov. Proxemic interaction in a multi-room music system. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, OzCHI '13, pages 153–162, New York, NY, USA, 2013. ACM.

[62] M. Strauch. Quick Sequence Diagram Editor. http://sdedit.sourceforge.net/index.html, 2014.

[63] N. Tanibata, N. Shimada, and Y. Shirai. Extraction of hand features for recognition of sign language words. In *International Conference on Vision Interface*, pages 391–398, 2002.

[64] I.K. Tran and S. Shah. Modeling motion of body parts for action recognition. In *Proceedings of the British Machine Vision Conference*, pages 1–12, 2011.

[65] R.D. Vatavu. User-defined gestures for free-hand TV control. In *Proceedings of the 10th European Conference on Interactive TV and Video*, pages 45–48, 2012.

[66] R.D. Vatavu. A comparative study of user-defined handheld vs. freehand gestures for home entertainment environments. *Journal of Ambient Intelligence and Smart Environments*, 5(2):187–211, 2013.

[67] R.D. Vatavu. There's a world outside your TV: Exploring interactions beyond the physical TV screen. In *Proceedings of the 11th european conference on Interactive TV and video*, pages 143–152, 2013.

[68] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision and Image Understanding*, 81(3):358–384, 2001.

[69] M. Wang, S. Boring, and S. Greenberg. Proxemic peddler: A public advertising display that captures and preserves the attention of a passerby. In *Proceedings of the 2012 International Symposium on Pervasive Displays*, PerDis '12, pages 3:1–3:6, New York, NY, USA, 2012. ACM.

[70] R. Want, A. Hopper, V. Falc ao, and J. Gibbons. The active badge location system. *ACM Transactions on Information Systems*, 10(1):91–102, Jan 1992.

[71] M. Weiser. The computer for the 21st century. *Scientific American*, pages 94–104, 1991.

[72] M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9(4):625–636, 2002.

[73] J.O. Wobbrock, H. Aung, B. Rothrock, and B.A. Myers. Maximizing the guessability of symbolic input. In *Proceedings of the CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1869–1872, 2005.

[74] J.O. Wobbrock, M.R. Morris, and A.D. Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1083–1092, 2009.

[75] J.J. Wu, R.E. Rink, T.M. Caelli, and V.G. Gourishankar. Recovery of the 3D location and motion of a rigid object through camera image (an extended Kalman filter approach). *International Journal of Computer Vision*, 2(4):373–394, 1989.

[76] M.H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074, 2002.

[77] E. Yoruk, E. Konukoglu, B. Sankur, and J. Darbon. Shape-based hand recognition. *IEEE Transactions on Image Processing*, 15(7):1803–1815, 2006.

[78] M.A. Youssef and A. Agrawala. Handling samples correlation in the horus system. In *INFOCOM 2004. 23rd Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 1023–1031. IEEE, 2004.

[79] M.A. Youssef, A. Agrawala, and A. Udaya Shankar. WLAN location determination via clustering and probability distributions. In *Pervasive Computing and Communications, (PerCom 2003). Proceedings of the First IEEE International Conference on*, pages 143–150. IEEE, 2003.

[80] Y. Zhu, B. Dariush, and K. Fujimura. Kinematic self retargeting: A framework for human pose estimation. *Computer Vision and Image Understanding*, 114(12):1362–1375, 2010.