GESTURE RECOGNITION OF EVERYDAY ACTIVITIES

By

KU NURUL FAZIRA KU AZIR
(902640)

MASTER THESIS

THESIS SUBMITTED IN PARTIAL FULFILMENT FOR
THE REQUIREMENTS OF THE DEGREE
MASTER OF SCIENCE IN HUMAN CENTERED SYSTEM
IN THE FACULTY OF
ELECTRONIC, ELECTRICAL AND COMPUTER ENGINEERING
OF THE UNIVERSITY OF BIRMINGHAM

SUPERVISOR:

DR. CHRIS BABER

University of Birmingham
Electronic, Electrical and Computer Engineering
Gisbert Kapp Building
Birmingham

August 2008

ABSTRACT

Multimodal recognition system is becoming a more common interaction tools in the fields of ubiquitous and wearable computing. Recent technologies and development of multimodal in human computer interaction have encouraged the notion and analysis of multimodal in human daily life activities.

This thesis explores the concept of multimodal i.e. speech and gesture recognition in everyday life activities. It propose an approach to recognize goal of activity based on detecting and analyzing sequence of gesture, speech, object, actions and locations that are being manipulated by the users. In domains such as cooking, where there are involve many similar and repeated of objects and actions can be a valuable and interest area to study in determining the concept of multimodal in everyday activities. An experiment of gesture and speech in cooking activity were analysed in term of object manipulation and sequence of actions by using video analysis and RFID tagged objects. There were compared with multimodal in computerized interaction.

This study has demonstrate multimodal also been used during cooking activity. Combination of speech and gesture results set sequence of actions which be used to determine the goal of activity through ontology multimodal. It also demonstrate a set of actions, objects and locations sequence guide to a new multimodalities in real life activities.

ACKNOWLEDGEMENT

I would like to acknowledge and extend my heartfelt gratitude to all those who gave me the possibility to complete this thesis. I want to thank the Department of Electronic, Electrical and Computer Engineering of The University of Birmingham, for giving this opportunity to complete this thesis.

I am deeply indebted to my supervisor Prof. Dr. Chris Baber whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. Special thanks go to James Cormley and CLIC Teams for helping and allowing me to use their hardware and software for this thesis. Without the assistance this thesis would not have been successful. Grateful thanks also to my tutor, Prof. Bob Stone whose keep give me the courage and support from behind.

I would like to thank my family members, especially my parents for supporting and encouraging me to pursue this degree. Without them, I would not have finished the degree. Thanks also to my colleagues in MSc program for giving support and encouragement through out our tough times together.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

The existence of multimodal on human-computer interaction (HCI) has been a subject of considerable interest since 1980s. Multimodal represents a new direction of computer interaction, ubiquitous and wearable computing. The system is shifting away from current GUI (graphical user interface) to PUI (perceptual user interface) (Turk, 1998, 30) with a design to support usability, efficiency, availability and robust system.

Multimodal interface system has developed rapidly after the introduction to "Put That There" system by Bolt (1980) [6] with steady progress toward building more system to suit the user's task, environment and user's need [11]. A common type of multimodal system is a combination of gesture and speech information. Multimodal applications range from map-based, military simulation, identification and verification to medical informatics and virtual reality.

In everyday life, both senses (gesture and speech) combine to give complementary and additional information about the world. The combination of these two senses gives much of the information about human daily everyday environment and activities. Such activities include are conversation, show direction, meeting, buying and selling and others.

Of central interest in this project is how gesture and speech in everyday life activities can be use as multimodal input. The approach involved the significant use of objects manipulation in cooking activity.

1.2 Cooking Domain

Much early work on the analysis of multimodal took place within the computer vision communities and focused on the representation and modeling actions. Thus, it required objects and people to be instrumented, and does not involved types of objects such as metallic objects, food items and objects that are very small. In contrast to these works, this project interested in domains where gesture, speech and object manipulation are being experiment as part of multimodal observation in human daily life activities.

In domain such as cooking, many activities involved similar actions with similar typed of objects. It does involve relatively number of repeated actions such as picking, chopping and pouring. The cooking domain involves a large number of objects across different typed of actions. The most important aspect, cooking is an activity that away from management task or computer interaction features. Through cooking, gesture can be identify through the object and speech that been manipulated. For example, Vybornova et. al (2007) [7] has suggest that when intending to perform an action, the user might

- Use words, to speak and describe an actions or intentions. Linguistic and action recognition are complementary.
- Doing action or intentions, but speaking something irrelevant to the actions that been perform. These two modalities will be analyzed separately, without merging.
- Users might contradict the action performed. Actions has priority, users might perform action first followed by speech.
- Just silent when performing actions.

Therefore, this project has conducted an extensive set of experiments on the classifications of kitchen activities carried out in two different typed of realistic settings. The experiments will be analyzed and conducted based on above probabilities of intention. It demonstrated how objects are being manipulated by users through gesture and assist by speech. The general idea of these

probabilities is to predict the next intention of user's movement when system know what object users interacting with and what action they are performing the object where at the same time to predict the goal of activity.

1.3 Thesis Aim

In this section the main aims of the thesis will be summarized. The overall aim of this research is to explore multimodal, i.e. speech and gesture recognition in 'real life' activity. By doing this, it is hoped that gesture and speech from cooking activity will be shown to be effective as one of multimodal input. The main aims of this project are:

- To investigate whether computerized multimodal results' are similar or not with the 'real world' multimodal.
- To predict next user's intention of action, activity and movement using ontology multimodal.
- To produce an intelligent environment based from result of second objectives.

1.4 Ontology Multimodal

The term 'ontology' came from field of philosophy, means systematic account of existence. In the context of computer and information science, ontology defines as a set of representational primitive with which to model a domain of knowledge or discourse [13]. Gruber (2008) suggest that representational primitive contains typical classes (or sets), attributes (or properties) and relationship (or relations among class members). This representational primitive represent a domain of action or knowledge. In the domain cooking, the representational primitive represent by object and gesture manipulate by users.

The representative and design of ontology is similar to tree diagram, in this project ontology use to predict the next intention of actions. A sequence of objects manipulation assumed to predict the next intention of actions and the

goal of activity. Once set of objects and sequence of actions been identified, a goal of activity can be identifying whether users making an omelette or egg fried rice. It is to believed also that speech, gesture and location recognition can be use to determine goal of activity in ontology if one of the elements is missing.

Generally, this thesis describes the multimodal interaction in real life activities. It specifically focuses on the gesture and object manipulation in a cooking domain. It will then discuss the multimodal architecture between the computerized and real life activities. The result of this analysis is then to predict the next intention of user's action and gesture through the development of ontology multimodal. Finally, this project will discuss the result of this project for the future development i.e. intelligent kitchen.

1.5 Intelligent Environment

Recently, there were some of researches on intelligent environment with multimodal interaction. For example by Jen et. al (2007) and Chi et al (2007) has promote a smart kitchen to promote healthy cooking. One of the purposes of this thesis is to create smart kitchen using multimodal input. This environment is more like a sensor-rich environment, by identifying movements of actions or object manipulation to assist users in kitchen.

Intelligent environment can be defined as a location (in this term, kitchen) that equipped with sensors, actuators and computers that are network with each other and internet (Le Gal, et al, 2001). Therefore it is hope through this experiment; system can be identifying the intention of users in kitchen. For example, identify the object manipulation when users are aim to make a chicken curry. Moreover, the aim of smart kitchen is to assist users i.e. elderly people or for youngster (first time in kitchen).

1.5 Contents of Thesis

The structure of this thesis follows firstly by understanding the multimodal and end with how analysis in day life activities can offer new perspective of multimodal.

This chapter explained the utilities and desired of using cooking as multimodal discovery in real life activity.

Chapter two gives an introduction to gesture and multimodal. It also provides a background of existing research in multimodal. Chapter three explains the motivations and analysis behind this project. Followed by chapter four, outline methods and materials of this project by perform a realistic cooking activity.

Chapter five shows result of gesture and speech recognition through cooking activity, such as making an omelette and egg fried rice. Chapter six illustrate a discussion and argument of gesture, object, actions, speech and locations recognition in cooking activity, hope to give new dimensions of multimodal recognition system, especially in human daily activity. This chapter will also discuss the future direction of this project and end this thesis with some conclusion on chapter seven.

CHAPTER 2

Literature Review

2.1 Overview of Chapter

This literature review starts with an introduction to the identifying and classification of gestures. Then, the discussion continues with speech and object manipulation.

The review then moves on to the review of gesture in human-computer interaction, and followed by a review of multimodal interaction systems. An evaluation of current design in practice is made and a case argued for focusing on identifying gesture recognition.

2.2 Gesture in the Everyday World

In a wide range of everyday life activities, human interactions have evolved a large variety of gestures. Gestures may involve the movements of body parts such as hands, arms, eyes, face or head. In fact every physical action involves gesture. Furthermore, the nature of gesture is an important component in human communication. This thesis will focus on the hand gestures especially on the motions of fingers, hands and arms. Therefore, this chapter will discuss the interaction of hand gesture in human daily activities and the implementation of gestures in the fields of ubiquitous and wearable computing such as multimodal.

There have been a consideration number of reports describing and defining gestures. Gestures have only been discussed in the field of linguistics but also

in the fields of computer vision and human-computer interaction for decades. Commonly, gestures have been described as a non-verbal communication usually involving movement of the body to express an idea, emotion, object or place.

According to Kurtenbach and Hulteen (1990) cited in Buxton (2002) [7] gesture can be defined as:

"…a motion of the body that contains information. Waving goodbye is a gesture. Pressing a key on a keyboard is not a gesture because the motion of a finger on its way to hitting a key is neither observed nor significant. All that matters are which key was pressed".

Therefore, from the view of Kurthenbach and Hulteen, gesture is a movement of the human body that signifies information. This is true; especially sign language, where gesture and pattern of fingers and hand convey individual meaning and information. Meaning refers to information contributes to specific goal.

From this, we could say that a gesture made by a person signify some purpose. This purpose could be defined by communication (as the examples in this paragraph imply), but could also be related to other actions. To return to the 'typing' example, while the act of pressing a specific key might not be a 'gesture' in this sense, the activity of typing could be recognized by a computer and used to interpret the person's current purpose.

For example, knowing that the person is engaged in rapid and prolonged typing might lead to the inference that the person is either busy or is engaged in a creative activity and should not be disturbed until there is a break in the activity. Table 2.1 shows form of gestures in human daily activities.

Table 2.1: Common Gestures in Human Daily Activities

ACTIVITY	GESTURES
Praying	Two hands flat together
Begging	Flats hand combine together
Conducting an orchestra	Varieties of gestures using hands
Martial arts	Variety movements of arms and body
Traffic control of cars and airplane	Hands flat pointing or moving
Everyday of conversations	Chatting, describe route or negotiating prices on market,

There is to believe that gestures are an addition or aid towards words spoken by human. It is to be true, as in daily activities human most likely to move hands or body while they speak without intend to do it. For example, people intend to move hands while speaking in public such as in daily conversation or public speaking competition. As Hummels & Steppers (2002) [7] argue, the mapping between hand movements and postures could lead to different meaning.

However, there is certain communications that can only be express using gestures such as sign language, i.e. American Sign Language (ASL). Any sign language has assigned meaning, strong rules of context and grammar may be applied to make recognition tractable.

Since then, researchers start to develop an ASL recognizer an application that concentrate on isolated signs or finger spelling by using instrumented glove or desktop based camera system. The recognizer used to help the physically impaired to interact with computer e.g. interpreting sign language.

Afterwards, researchers believe the concepts of sign language recognizer might also be used in developing of advance human computer interactions (HCI). According to Rodney Brooks, director of MIT's Artificial Intelligent Lab; though researchers had made great progress in improving verbal

communication between computer and human, a progress of non-verbal communication had made doubles of progress and success. As a result, traditional HCI had been move towards advance and maturity interactions based on gesture recognition.

Gesture recognition involves recognizing and interpreting the physical movements made by humans and this typically involves mathematical algorithms and techniques from computer vision and image processing. As Je et al. report, gesture recognition is human interaction with a machine, i.e., a computer, in which human gestures are recognized by the machine (Je et al., 2007). The use of gesture recognition for human-computer communication is related closely to development of intelligent human machine interaction. As Westeryn shows, gesture recognition can be considered as a part of pattern recognition and is becoming a common interaction tool in the fields of ubiquitous and wearable computing (Westeryn et al., 2003).

Since the 1970s, developmental researchers have investigated the link between gesture and human computer interaction. Numerous works had been done regarding gesture recognition [22] – [24]. These typed of application applied gesture recognition as an input device which uses body movement i.e. eye or hand as input control device or application instead of joystick, mouse or keyboard. As Je et al (2007) and Mitra et al (2007) indicate hand gestures can provide meaningful information of human intention and interacting with the environment. Gesture recognition involves a range of applications from sign language recognition through medical rehabilitation to virtual reality.

Although hundreds of works and researches had been done in recognizing gesture in computer interactions, there is less works has been done in recognizing gesture in daily life activities. For that reason, this thesis will concentrate on exploring gesture and multimodal in daily life activities before move to gesture based computer interactions such as intelligent environment. Thus, it is essential to understanding the clarification of gesture first.

Hence, the following sections will review a psychology and anthropology literature of classifications of gestures in daily life activities and human computer interaction.

2.3 Classifications of Gesture

Gestures are used for many activities and application involving different part of body movements. This section will review classifications of gestures focusing on hand gestures as related to incorporating the understanding of gesture in human daily activities. In addition, a classification of hand gesture in human computer interactions also will review.

Previously, movements of gestures and meaningful interactions are quite complex to be understands. There are growing body of literature studying on human gesture and classifications of gesture. Numbers of authors had made different perspective to classify gestures. According to Billinghurst & Buxton (2002) [7], gesture may exist in isolation or involve external objects. This suggests that gestures can be classified according to their function and that they are not simply a matter of communicating information (as Kurtenbach and Hulteen (1990) implied). As Cadoz (1994) proposed, he classifies hand gestures according to their functions. Table 2.2 shows classifications of gestures proposed by Cadoz.

Table 2.2: Classification of Gesture proposed by Cadoz (2004)

Classifications	Purpose
Semiotic	Used to communicate meaningful information and results shared from cultural experience
Ergotic	Used to manipulate the physicals world and create artifacts
Epistemic	Allows human to learn from environment through tactile experiments and haptic exploration

As indicate on above table, semiotic gesture connected with speech while ergotic gesture manipulate with physical object in real life. The use of hand gesture can manipulate objects by changing the object position, orientation and shape.

It is immediately clear from the above discussion; gestures can be divided into two major groups. One group involves gesture with communication (empty handed gesture), and another involving manipulation and gasping of objects. As Billinghurst points out, a set of gestures which relate to object manipulation which these range from pointing to objects, touching or moving objects, changing the objects' shape, activating objects such as controls or handing object to others (Billinghurst, 2002). This statement also had been proven by McNeill (2000) which demonstrate pointing is one of way people display their knowledge to indicate place, object or object move from one place to another.

2.3.1 Gesture and Speech

Although gesture is not equivalent to speech, but gesture and speech is complement to each other. Many researchers believed in this statement as McNeill (1992) reveal that speech and gesture are an integrated from of expressions of utterances where speech and gestures are complementary Gesture accompanied by speech also called gesticulations. Gesticulations form 90% of everyday used gestures. McNeill (1992) cited in Hummels & Steppers (2002) [7] points out spatial information that is hard to capture can be conveyed by gesture.

Kendon (1988) cited in Mulder (1996) proposed four (4) types of relationship between gesture and speech which been summarize in Table 2.3. This work also been proposed by McNeill (1992).

Table 2.3: Classifications of Gesture and Speech

Types of Gestures	Explanation
Gesticulation	Ranging of pointing objects using one's hands to provide emphasis when speaking. Spontaneous movement of hand and arms during speech.
Pantomime	A significant gesture without speech, used in theater to communicate a story. For example, a dumb show.
Emblems	Gesture that have single meaning to individual that perform it. For example, "OK" or "V" gestures made from forming the fingers.
Sign Language	A defined set of gestures that correspond to words and letters to fulfill linguistic communication system.

Source: Mulder (1996); *Hand Gesture for HCI*

In addition to Kendon's taxonomy, McNeill (1992) proposed another type of gesture classifications. This type of classifications is related to the process communications which known as *beat* and *cohesive* gesture (McNeill, 1992 cited in Billingburst & Buxton, 2002). Beat gestures related to the movement of hand up and down with the rhythm of speech e.g. in debate competition. While cohesive gesture related to variations of iconic, pantomimic and deitic gesture and related to portions of discourse. Iconic, pantomimic and deitic gesture will be review later in this section.

McNeill also suggested that gesture can be considered from two different approaches. One approach views gesture as the individual speaker-gesturer's ongoing mental life (and so recognizing gesture might allow us to infer what the person is thinking or planning to do), while the other approach view gesture as part of social interaction which the person participates (and so recognizing gesture might allow us to better interpret the communication of that person).

However it is important to bear in mind that the use of gesture might not produce smooth speech but create disfluency and delay the speech. This

observation is supported by Rauscher et al (1996) whose mention that if gesture is inhibited, it can also affect the fluency of speech and increase the proportion of time spend pausing (Graham & Heywood, 1975).

Though there are several more classifications of gesture with speech, it is clearly shows that Kendon and McNeill's classification more recognizable and stresses the strong relationship with gesture and speech. Furthermore, each of the above categories of gesture had been also been identify in human computer interaction which will be discussed later in this chapter.

2.3.2 Gesture and Object Manipulation

As been discussed in previous section, another major group of gesture is relationship of gesture and object manipulation. This type of relationship related to ergotic gesture. However a research on gesture and object manipulation relationship is quite limited compare to gesture and speech especially in recognizing of objects in daily life activities. In this section, a review of gesture and object manipulation will be based on three (3) of classification made by different authors.

However, it is essential to look on different types of gesture with object manipulation first. According to Mulder (1996), gesture and object manipulation can be divided into few groups, there are:

- Goal directed manipulation
    - Changing position of an object
    - Changing orientation of an object
    - Changing shape of an object
    - Contact with an object
    - Joining objects
    - Indirect manipulation
- Empty handed gestures
- Haptic Exploration

Appendix A list actions that involve gesture and object manipulation in group of goal directed manipulation, empty handed gesture and haptic exploration. The appendix shows some of movement or actions that involve common gesture in human daily life activities (see Appendix A).

There had been argues by researchers in classification of gesture based on object manipulation. Some of researchers believe ergotic gestures should also been classified as similar to semiotic gesture while there are also researchers believe the classification should based on the physical appearance of object that involved.

First classification that will be review is based on the physical characteristic of an object. Mulder (1996) believes since human have a power and intention to change the position, orientation and shape of an object therefore classification of gesture and object manipulation should based on object's characteristic. Table 2.4 shows classifications of gesture and object manipulation proposed by Mulder (1996).

Table 2.4: Classifications of Ergotic Gesture based on characteristic of Object

Characteristics	Actions
Object typed	Solid, fluid or gaseous
Change effectuation	Position, Orientation or shape
Hand involved	One or two
Indirection level	Direct manipulation or through other object or tool

However, the value of above classification is not be recognized and limited due to the lack and no reference had made to the task of hand [23]. This typed of classifications is more concern the cognitive elements and process in hand gestures and movements.

Thus, an appropriate ways to classified gesture with object manipulation is based on their functions; similar to semiotic gesture. Mulder (1996) indicate the classification can be classified as *prehensile* and *non-prehensile.* Such non-prehensile movement are pushing, lifting, tapping and punching. Mackenzie (1994) cited in Mulder (1996) defined prehensile gesture as:

"*The application of functionality effective forces by the hand to an object for a task, given numerous constraints"*

One of recognizable taxonomies that identified prehensile movement had been proposed by Napier (1993) cited in Mulder (1996). Napier reveals this type of taxonomy can be either: -

- Precision grip
- Power grip
- Hook grip
- Scissor grip

Napier demonstrated that this type of taxonomy relates to muscle skeletal properties of hand and notable oppositions which it incorporate with notion of task and required precision or power (Napier, 1993 cited in Mulder, 1996). However, neither scissor or hook grip does not relate to the notion of task. The grip more likely refer to the frequently movement of hand been used and the function of the grip itself refer to the activities that been perform not the physical characteristic of an object used. Therefore, this classification is an ambiguous.

Then Pressing (1999) cited in Mulder (1996) had also proposed taxonomy to classified ergotic hand movements. Pressings demonstrate this type of taxonomy in three (3) categories. There are:

- Use of control effect : i.e. modulation (parametric change), selection (discrete change) or excitation (input energy)
- Use of kinetic image: i.e. scrape, slide, ruffle, crunch, glide, caress etc.

- Use of spatial trajectory: i.e. up, down, left, right, in, out, circular, sinusoidal, and etc.

In generally, these three classifications have own limitation. None of it had made the exact classification of gesture and object manipulation relationship. In Mulder taxonomy, the classification based on the control task taxonomy for specific purpose while Napier's taxonomy emphasis the observers point of view and extract semiotic function which demonstrate semiotic function always present. In a mean time, Presssing's taxonomy based on the shape of hand as parameters and has same limitations to Mulder's taxonomy.

In a mean time, even there are no precise classifications of gesture and object manipulation, there had been numerous work on finding an interaction with everyday object in everyday life activities. These applications combine the affordances and richness of everyday objects interaction with the power and richness of digital world (Feldman et al, 2005).

On of the research design is ReachMedia application proposed by Feldman et al (2005) [30]. The ReachMedia application performs by applying wristband approach. The use of RFId tag and readers is to provide an information and detect an objects that the uses interacting with. The detection process involved "interaction on the move" whish users engaged with primary task other than interaction itself.

Another work involved object and gesture manipulation is detecting arm gesture of eating and drinking activity proposed by AmFit, O. et al (2005). This research also involves body-worn sensors to detect gestures that related to food intake. The outcome of this research is to provide information and monitoring of user's diet by looking at the hand gesture.

From the above review and discussion, it is clearly that there is no exact classification of ergotic gesture that been made. Therefore, in this thesis a combination each of classification will be used to recognize the relationship of

gesture and object manipulation in both real life activities and human computer interaction.

2.4 Gesture in Human Computer Interaction

In the Human Computer Interaction (HCI) literature, the word gesture has been used to identify hand movement to control computer process and application. As been discussed from previous section, gesture in an important part of human conversional interaction. Gesture had been studied broadly in recent years in effort to build HCI interfaces that go beyond conventional input devices such as keyboard and mouse manipulations.

The gesture in HCI not only involved pointing but also hand shape as an input devices or known as *hand centered input* [36]. The use of hand as a communication device can be applied in both receiving and sending information. Starman (1993) defined gesture in HCI as:

> *"Whole hand input as the full and direct use of the hand's capabilities for the control of computer-mediated tasks".*

This statement also been agreed by Je et al (2007) and Mitra et al (2007) which indicate hand gesture can provide meaningful information intention and interacting with the environment. Hummels & Stepper (2002) also argues that most currently design of HCI involved set of gestures and simple movement, which required users to learn.

This sections, presents a review of classification of gesture in HCI. Later, an overview of some application based on gesture recognition.

2.4.1 Classification of gesture in HCI

Since the development of gesture in HCI had been growth rapidly in 1980s, there had been hundreds of an application design based on gesture recognition. The design may imply semiotic, ergotic gesture or both.