Bring Your Body into Action

Body Gesture Detection, Tracking, and Analysis for Natural Interaction

FARID ABEDAN KONDORI

Doctoral Thesis
Department of Applied Physics and Electronics
Umeå University
Umeå, Sweden 2014
Abstract

Due to the large influx of computers in our daily lives, human-computer interaction has become crucially important. For a long time, focusing on what users need has been critical for designing interaction methods. However, new perspective tends to extend this attitude to encompass how human desires, interests, and ambitions can be met and supported. This implies that the way we interact with computers should be revisited. Centralizing human values rather than user needs is of the utmost importance for providing new interaction techniques. These values drive our decisions and actions, and are essential to what makes us human. This motivated us to introduce new interaction methods that will support human values, particularly human well-being.

The aim of this thesis is to design new interaction methods that will empower human to have a healthy, intuitive, and pleasurable interaction with tomorrow’s digital world. In order to achieve this aim, this research is concerned with developing theories and techniques for exploring interaction methods beyond keyboard and mouse, utilizing human body. Therefore, this thesis addresses a very fundamental problem, human motion analysis. Technical contributions of this thesis introduce computer vision-based, marker-less systems to estimate and analyze body motion. The main focus of this research work is on head and hand motion analysis due to the fact that they are the most frequently used body parts for interacting with computers. This thesis gives an insight into the technical challenges and provides new perspectives and robust techniques for solving the problem.

Keywords: Human Well-Being, Bodily Interaction, Natural Interaction, Human Motion Analysis, Active Motion Estimation, Direct Motion Estimation, Head Pose Estimation, Hand Pose Estimation.
Acknowledgments

I am grateful to the following people who have directly or indirectly contributed to the work in this thesis and deserve acknowledgment.

First of all, I would like to thank my supervisors Assoc. Prof. Li Liu and Prof. Haibo Li for not only providing this research opportunity, but also for the positive environment they created in our research group. I am truly indebted and thankful for the valuable guidance from them in the research world. Thanks for all the inspirations and encouragements during these years. Without your help and support I hadn’t been able to pursue my research.

I would also like to thank my best friend, and my best colleague, Shahrouz Yousefi. We had many collaborations, interesting discussions and enjoyable moments. Special gratitude goes to my friends and colleagues, Jean-Paul Kouma, Shafiq Ur Rehman, Alaa Halawani, and Ulrik Söderström for their helpful suggestions and discussions. Thanks to all the staff of the department of Applied Physics and Electronics (TFE) for creating an enjoyable and pleasant working environment. I would like to thank the Center for Medical Technology and Physics (CMTF) at Umeå University for financing part of my research.

Most importantly, I owe sincere and earnest thankfulness to my parents and my brothers, without whom none of this would be possible, for all the love and support they provide. Finally, I would especially like to thank my dear fiancée, Zeynab. Thanks for bringing endless peace and happiness to my life.

Thank you all.

Farid Abedan Kondori
Umeå, June 2014
Contents

Abstract i
Acknowledgments iii

1 Introduction 1
   1.1 Aim of the thesis ........................................ 2
   1.2 Research problem ....................................... 2
   1.3 Thesis outline .......................................... 3

I WHY 5

2 Design for Values 7
   2.1 From user experience to human values ................. 7
   2.2 Human well-being ...................................... 8

3 Human Engagement with Technology 11
   3.1 Integrating surgery with imaging technology ....... 11
   3.2 From clinical labs to individual homes ............... 12

4 The Growth of Technology 15
   4.1 Mobile computers ..................................... 15
   4.2 Wearable computers .................................... 16
   4.3 New sensors ............................................ 18

5 Introducing New Interaction Methods 21
Chapter 1

Introduction

The emergence of computers has reshaped our world. From late 20th century, when there was one computer for thousands of users, to 21st century where there are thousands of computers for one user, technology has changed the way we grow up, live together, and grow older.

Our interaction with computerized devices has also changed. The shift from typing commands as the only input channel, to running applications by just clicking on icons in graphical user interfaces was the major step in the evolution of human-computer interaction. As digital technologies advanced, so there has been a turn towards touch-based interface systems, where users perform different hand gestures to operate their devices.

In the near future, computers will continue to proliferate, from inside our bodies to roaming Mars. They will also look quite different from the PCs, laptops, and handheld electronic gadgets of today. Now the question is: what should our interaction with computers be like in the coming years? Will current methods enable us to have enjoyable life with computers? And if not, what kind of interfaces would be more useful to improve quality of human life in the future?

Obviously existing interaction methods will face a lot of challenges, due largely to the fact that they are designed to satisfy current user needs. However, in the next decade, or in the next few years, users will demand more from computers, and technology will affect all aspect of human life. Therefore, in order to explore interaction methods for the future world, we need to anticipate future human needs. To do so, detailed observation of human engagement with digital technologies will be of great importance.
Nevertheless, good understanding of human values, and centralizing them in
the design and development process, will be the key factor to succeed. And of
course, all of this is possible when we are provided with recent technologies.
Consequently, to design and develop new interaction techniques, three key ar-
eas must be taken into account: human values, human engagement, and new
technologies (Fig. 1.1).

1.1 Aim of the thesis

The aim of this thesis is to understand, study, and analyze the three aforemen-
tioned factors, and provide new interaction methods that will empower human
to have a healthy, intuitive, and pleasurable interaction with tomorrow’s digi-
tal world. A natural and intuitive interaction enables users to operate through
intuitive actions related to natural, everyday human behavior.

1.2 Research problem

In order to achieve the aim, this research is concerned with developing theories
and techniques for exploring interaction methods beyond keyboard and mouse,
utilizing human body. To accomplish this, we address a very fundamental problem, *human motion analysis*. More specifically, this thesis attempts to develop vision-based, marker-less systems for analyzing head and hand motion.

1.3 Thesis outline

The thesis is carefully organized in four parts. In order to help readers to follow the thesis smoothly, three principle questions will be answered throughout the thesis:

1. *Why this research is done?*

2. *What is done to achieve the goal?*

3. *How the research goal is achieved?*

Part I provides a concrete reason that *why* this thesis is done. Three main aspects will be addressed, and the need to introduce new interaction methods will be discussed.

*What* we have done to achieve our goal is explained in Part II. In this part we will define a scientific problem on how to provide new interaction methods and address the technical challenges.

Third part will discuss *how* we have accomplished our goal. New perspectives and robust techniques for solving the problem are presented in this part. Then, contributions, concluding remarks, and future directions will be highlighted.

Eventually, all publications that provide basis for this thesis are listed in the last part.
Part I

WHY
Chapter 2

Design for Values

Computers has reshaped our world and changed the way we live. New generation of technologies is providing us with more creative uses of computing than ever before. Computers will continue to proliferate, and affect all aspects of our lives in tomorrow’s digital world. They will also look quite different from computerized devices of today. The world we live in is changing by new technologies, and the extent of these changes will increase in the near future.

However, transformations in computing technologies are not the main concerns about the future world. The main issue is how technologies can support the things that matter to people in their lives, human values. These values drive our decisions and actions, and are essential to what makes us human. Being active and healthy, self-expression, creativity, and taking care of loved ones are such values. When considering the future digital world, we should be aware that the attributes that make us human should be manifest in our interaction with technology.

2.1 From user experience to human values

For a long time, what has been crucial for human-computer interaction (HCI) researchers was focusing on user-centered design, centralizing what users need and want from technology. Nevertheless, new perspective in HCI tends to extend this attitude to encompass how human desires, interests, and ambitions can be met and supported through technology.
Currently, technologies do not necessarily solve users’ problems as they used to, but increasingly are enabling users to achieve other kinds of interests and aspirations. For instance, Instagram, a photo and video sharing application, is designed specifically to support certain values, such as enabling people to express themselves, to show their affection to others, and to demonstrate their creativity (Fig. 2.1).

In short, recognizing human values and centralizing them in the design and developing of interaction methods is critical.

### 2.2 Human well-being

Human well-being is one of the core values that should be recognized and supported by the future technology [1]. The motivation for this work is that being active is central to human well-being in many situations and over time [2]. Indeed, movement is part of the essence of being human. The first abilities
developed by humans were physical skills. Our ancestors learned how to use their hands to make tools for hunting animals, protecting their families, and building their communities. We, as human beings, move. From early childhood to our last step, we are designed to move.

When computer technology grew into our workplaces and other areas of our everyday life, many of the opportunities to use our physical abilities diminished. We have utilized computers in many situations to relieve ourselves from physically demanding tasks. For many years, the main physical interaction between humans and computer technology has been our fingertips on the keyboard and some clicking with the mouse while the rest of our body has remained seated. Passive interaction with technology changed the macro-monotony of human movements into the micro-monotony in front of computers.

We need to remember that physical activity is fundamental to human well-being. It has long been known that regular exercise is good for our physical health. It can reduce the risk of cancer, heart disease and strokes. In recent years, studies have shown that regular physical activity also has benefits for our mental and emotional health [3]. Emotional well-being and mental health concerns can dramatically affect physical health. Stress, depression, and anxiety can contribute to a host of physical ailments including digestive disorders, sleep disturbances, and lack of energy. Exercise can help people with depression and prevent them becoming depressed in the first place.

Physical, mental, and emotional health can improve through a balanced level of physical activity in everyday life. It has turned out that two-thirds of the adult populations (people aged 15 years or more) in the European Union do not achieve recommended levels of activity [4]. We need to move, otherwise we risk developing physical and mental health problems.

Another aspect of human well-being encompasses our interpersonal relationships, social skills, and community engagement. At the soul of every human is the need to feel a sense of togetherness. This is an extension of our dependency and our care for our fellow human beings. From early childhood we are taught caring for friends, family members, neighbors and relatives which leads to a need for togetherness and helps build community. This can be seen when we come together at festivals or any social ritual - there is always this sense of community. When we do not have this, we feel alone with no one to talk to or share with. In the higher level, this can affect the unity in societies. We need to be together to build and develop our societies.
In our digital world, we are surrounded by computers. We are spending a great deal of time working with laptops, smartphones, and tablets in our daily lives. With the emergence of teleworking, today we do not even need to physically interact with our employers and customers. Many people stay at home and work remotely. This increasing tendency has led to isolation of individuals in our societies. Despite using technology to communicate with others, many people feel alone. Although social networks have made it easier to connect to friends, still loneliness is evident. There is a classical belief that technology can be employed to remedy loneliness, to connect people to each other, but today we feel more alone. Based on the study conducted in Australia, the more we use technology to connect to others, the more we feel alone [5].

Social networks and communication technologies are meant to connect us, but why they are making us feel alone? Answer to this question is beyond the scope of this thesis; however there is a clue that might help. Research reports reveal that only 20 to 40 percent of interpersonal communications involves explicit meaning of words and verbal cues, while 60 to 80 percent is conveyed through non-verbal signals [6]. When we communicate, non-verbal cues can be as important as, or in some cases even more important than, what we say. Non-verbal communications include tone of the voice, facial expressions, and body language or body movements. Obviously using body movements have a great impact when we communicate, the fact that is missing when we interact with technology.

Passive interaction with social networks might be one of the reasons that make us feel lonelier. Imagine a bodily active social network where users interact with and communicate with each other through body movements. Active interaction through body movements can convey closeness and togetherness, the things we need as human beings in our lives. Perhaps new generation of social networks will incorporate body movements to deliver more humanistic experiences.
Chapter 3

Human Engagement with Technology

To design our interaction with tomorrow’s digital world, a comprehensive study of today’s human engagement with technology is crucial. Not only will consideration of human involvement with digital technologies expose limitation of existing interaction methods, but it will provide us with vital clues to what extent future technology will empower human.

Several areas can be reviewed, however, we will consider two creative human engagements with digital technologies in medical applications.

3.1 Integrating surgery with imaging technology

Reviewing medical images and records of patients by surgeons is one of the crucial steps during surgery. Surgeons have to gather and analyze additional information on the state of a patient in surgery. Incorporating computers in the operating rooms enables a wide range of patient related data to be collected and analyzed. However, the optimal viewing of this information becomes problematic due to the sterility restriction in the operating rooms.

Since maintaining a pristine sterile field is substantially important in the operating rooms, using non-sterile objects to interact with computers cause a problem for surgeons. For instance, a surgeon may be required to step out of the surgical filed to operate the computer with keyboard and mouse to view MRI or CT images. The surgeon then has to scrub up to the elbows once again, and then re-enter the surgical field. Due to the time involved in re-
Figure 3.1: A modern operating room equipped with computers and imaging technology [7].

scrubbing, surgeons may not use computers as often as needed, and rely on the data from the last observations. This problem can be addressed by developing new interaction systems beyond the keyboard and mouse, utilizing natural hand gestures. Surgeons can control computers and manipulate screen contents simply by moving hands in the air, without ever having to sacrifice sterile filed. The need for new interaction systems is more evident with the emergence of robotic nurses in the tomorrow’s operating rooms [8]. The growth of techno-dependency is a common feature in the future operating rooms. As a result, designing natural and easy to learn interaction methods that meet the required sterility condition in the operating rooms will be of great importance.

3.2 From clinical labs to individual homes

Until recently, patients had to visit their therapists to monitor and evaluate serious health issues, movement disorders, for instance. Clinicians and physicians could monitor patient’s improvement after surgical interventions by tracking
Figure 3.2: Using rehabilitation gaming system for home-based therapy [9].

and analyzing patient’s movements. Clinical gait analysis labs are the most known existing technology for this task. Several markers and identifiers are mounted on the human body, which are detected and tracked by special sensors to monitor and analyze patient’s motion.

But healthcare systems are changing fast. Technological progress in telecommunications has led to the development of telerehabilitation systems for home-based therapy. Telerehabilitation offers systems to control rehabilitation process at distance. Communication via the Internet between patients and therapists enables telerehabilitation. Therapists can observe patients’ status and send commands via the Internet; thus patients do not need to visit rehabilitation facility for every rehabilitation session.

Telerehabilitation systems provide the possibility of early diagnosis, increase of the intensity and the duration of rehabilitation, start of therapy under acute conditions, and shortening of in-hospital period [10]. However, the change in healthcare systems is mainly due to the effectiveness of home setting in rehabilitation [10], i.e. taking patient’s treatment and observation process from clinical labs out to daily life activities. One of the drawbacks associated with clinical gait analysis labs is the patient’s unnatural behavior. It has turned out that patients are biased towards their physicians [11]. When it
is time for a follow-up check, patients might be affected or even intimidated by the specific circumstances in the lab environment. Therefore, they may not present their natural movement pattern, and try to present the best possible performance, pretending that everything is under control, though this is not true. In fact, the best way to assess patient’s improvement is to observe the patient in the daily life activities. Not only physical treatment process will be altered, but the way movement disorders are diagnosed will be also changed.

There will be a shift towards ubiquitous diagnostic systems, where computers will be utilized to analyze human daily life activities, literally everywhere. Humans will be monitored as they work, walk, play etc. This new perspective calls for intelligent diagnostic systems, where every human movement should be observed and analyzed to distinguish abnormal movement patterns.
Chapter 4

The Growth of Technology

In this chapter the rapid development of digital technologies and visual sensors, and the provided opportunity to design, develop, and implement new interaction techniques will be addressed.

4.1 Mobile computers

Since their inceptions, mobile phones have gone through a lot of changes. Without doubt, the first stage of mobile phone evolution was reducing the size. Users could proudly show their cellphone to other people, and then, put it back into their pocket. Adding colorful display and digital camera made cellphones look much better, and by then, users started to realize that cellphones can enable them to experience new functionalities, such as capturing and recording life events, and sending/receiving multimedia messages.

Nevertheless, Apple’s revolutionary product iPhone completely changed the concept of cellphones, and introduced new terminology, smartphone. iPhone’s touch display let users to have large visualization space as they use their fingers to open and run different applications. Equipped with high quality camera, and strong processor, it was able to run various applications, from map/navigation to different game applications. Apple’s success in the market made other competitors to manufacture similar products. And now, thanks to the high competition in the market, smartphones are of acceptable price, while equipped with high resolution cameras, and strong CPUs.

Emergence of smartphones was largely because of invention of touch screens. Replacing physical buttons with multi-touch displays provided new input chan-
Figure 4.1: Moving from 2D touch screen interaction space towards 3D space behind smartphones provides a great opportunity to design and develop natural user interfaces.

nel for users, and it was surprisingly successful. Although designers and manufacturers have explored various nice features of smartphones, they are still unable to utilize smartphone capacities to introduce new means of interaction. Since smartphones have high resolution camera and powerful CPUs, they provide us opportunities to employ computer vision algorithms for developing new interaction methods, to deliver new experiences (Fig. 4.1).

4.2 Wearable computers

Wearable computers are electronic devices that are worn by the bearer under, with or on top of clothing [12]. This class of wearable technology has been developed for general or special purpose information technologies and media development. One of the most known applications of wearable computers is augmented reality [12]. Augmented reality means to superimpose an extra
layer on a real world environment. The extra layer’s elements are computer generated inputs such as sound, graphics, GPS data etc. that let users receive more information on top of the real world view. Head-mounted displays are such wearable computers.

Recently Google launched their optical head-mounted display, Google Glass [13] for certain group of developers. Google Glass is operated through a touchpad and voice commands, and displays information in a smartphone-like hands-free format. Google Glass perfectly shows us the potential of future wearable computers, whose hardware include powerful processing unit, Internet connectivity, high resolution camera, display, and motion capture sensors. And what else do we need to create a totally new interaction method for this kind of future devices (Fig. 4.2).
4.3 New sensors

New visual sensors have opened a new angle for researchers to make use of image analysis algorithms in order to design new interaction techniques. Commercial high resolution cameras are becoming smaller, and therefore, can be conveniently mounted on the human body to analyze human activities. They are also capable of recording human daily life activities for further analysis.

More recently, sensing manufacturers have attempted to add one more dimension to conventional 2D RGB cameras. Creative Senz3D [14], Microsoft Kinect [15], Structure [16], and Leap Motion [17], are such sensors (Fig. 4.3). Kinect, for example, provides a robust solution to infer 3D scene information from a continuously-projected infrared structured light.

Growing availability of 3D sensors reflects the trend towards usage of 3D information to create innovative applications never experienced before. 3D sensors are integrated into smartphones and tablets quite recently, opening completely new world for creation of novel experiences (Fig. 4.4 and 4.5).
Figure 4.4: Structure sensor simply mounts around the side of the iPad, turning the tablet into a portable 3D scanner [18].

Figure 4.5: Google’s Project Tango, an Android prototype phone with 3D sensor [19].
Chapter 5

Introducing New Interaction Methods

This chapter will answer the first question; *Why this research is done?* For this reason, motivation, need, and technical feasibility will be addressed (Fig. 5.1).

- **Motivation**

  In the second chapter the importance of human values in the tomorrow’s digital world was emphasized. Centralizing human values rather than user needs in the future interaction systems is essential. Human well-being is one of the core values that should be supported through the future technology. The motivation for this work is to design and develop bodily interaction methods to achieve human well-being.

- **Need**

  Human involvement with new technologies is growing rapidly. The new generation of technologies is providing us with more creative uses of computing than ever before. The growth of human engagement illustrates to what extent future technology will empower human. Computers will definitely be employed for all kinds of novel engagements. A variety of people will use computing technologies, from children to the elderly, and from non-computer literate to professionals. Although rapid advancements of computing science will make all of these engagements possible, there is a need to introduce natural interaction methods to facilitate future human engagements.
• Feasibility

Digital technology is rapidly becoming inexpensive, everyday commodities. Mobile phones have evolved from handsets to the world in our hands. Wearable computers are no longer cumbersome, but are vastly considered as fashionable accessories. The world we live in is changing by new technologies, and the extent of these changes will increase in the near future. All of this means that the way we interact with the future technologies can be changed. Recent advancement in sensing technologies is strong evidence that we are provided with great opportunities.

Now it should be easy to answer why this work is done:

"To introduce new interaction methods that will empower human to have a healthy, intuitive, and pleasurable interaction with tomorrow’s digital world."
Part II

WHAT
Chapter 6

Interaction through Human Body

This chapter will discuss what needs to be done to accomplish the aim of the thesis.

6.1 Problem definition

In order to provide a healthy, intuitive, and pleasurable interaction, this research is concerned with developing theories and techniques for exploring interaction methods beyond keyboard and mouse, utilizing human body. In this thesis, human body is regarded as functional surface to interact with computers. It will be shown that human body position, orientation, and 3D motion can be utilized to facilitate natural and immersive interaction with computers. Therefore, to enable new interaction methods, we address a very fundamental problem, human motion analysis.

Human body motion can be described in terms of global rotation and translation $G(R,T)$, and local motion $L(\beta_1, \beta_2, \cdots, \beta_n)$. Thus, we formulate body motion as a function of two variables G and L:

$$Body \ motion = M(G(R,T), L(\beta_1, \beta_2, \cdots, \beta_n))$$

Considering the head, global motion $G(R,T)$ describes the six degree-of-freedom (DOF) head motion with respect to the real world coordinate system, while $L(\beta_1, \beta_2, \cdots, \beta_n)$ can represent lip motion, eye motion etc. The fo-
The focus of the thesis is to estimate global body motion to provide new interaction methods.

### 6.2 Human motion analysis

The study of human motion dates back to 1870s when Muybridge [20] started his work. Since then, the field of human motion analysis has grown in many directions. However, research and results that involve the recovery of human motion is still far from being satisfactory. The science of human motion analysis is fascinating because of its highly interdisciplinary nature and wide range of applications. The modeling, tracking, and understanding of human motion has gained more and more attention particularly in the last decade with the emergence of applications in sports sciences, human-machine interaction, medicine, biomechanics, entertainment, surveillance etc.

Fig. 6.1 illustrates different approaches to human motion analysis. Existing human motion tracking and analysis systems can be divided into two main groups: non-vision-based and vision-based systems.
6.2.1 Non-vision-based systems

In these systems, non-vision-based sensors are mounted on the human body in order to collect motion information and detect changes in body position. Several different types of sensors have been considered. Inertial and magnetic sensors are examples of widely used sensor types. Well-known types of inertial sensors are accelerometers and gyroscopes. An accelerometer is a device used to measure physical acceleration experienced by the user [21] [22]. One of the drawbacks of accelerometers is the lack of information about the rotation around the global Z-axis [23]. Hence, Gyroscopes, which are capable of measuring angular velocity, can be used in combination with accelerometers in order to give a complete description of orientation [24]. However, the major disadvantage of inertial sensors is the drift problem. New positions are calculated based on previous positions, meaning that any error in the measurements will be accumulated over time.

In [25], authors show that adaptive filtering techniques can be used to handle the drift problem. They also suggest that utilizing magnetometers can minimize the drift in some situations. The use of magnetometers, or magnetic sensors, is reported in several works [26] [27]. A magnetometer is a device that is used to measure the strength and direction of a magnetic field. The performance of magnetometers is affected by the availability of ferromagnetic materials in the surrounding environment [28].

6.2.2 Vision-based systems

Vision-based motion capture systems rely on cameras as optical sensors. Two different types can be identified: Marker-based and marker-less systems.

The idea behind marker-based systems is to place some type of visual identifiers on the joints to be tracked. Stereo cameras are then used to detect these markers and estimate the motion between consecutive frames. These systems are accurate and have been used successfully in biomedical applications [29] [30] [31]. Though, many difficulties are associated with such a configuration. For instance, scale changes (user distance from camera) and light conditions will seriously affect the performance. Additionally, marker-based systems suffer from occlusion (line of sight) problems whenever a required light path is blocked. Interference from other light sources or reflections may also be a problem, which can result in so-called ghost markers. The most important limitation for such systems is a need to use special markers to attach
to the human body; furthermore, human motion can only be analyzed in a predefined area covered by fixed, expensive cameras.

Marker-less systems try to employ computer vision algorithms to estimate human motion without using markers. The use of cheap cameras is possible in such systems [32] [33] [34]. Nevertheless, the markers removal comes with the price of complicating the estimation process of 3D non-rigid human motion. These systems can be divided into two groups: model-based and appearance-based systems.

**Model-based methods**

Model-based approaches use human body models and evaluate them on the available visual cues [35] [36] [37] [38] [39] [40] [41] [42] [43]. A matching function between visual input and the generated appearance of the human body model is needed to evaluate how well the model instantiation explains the visual input. This is performed by formulating an optimization problem whose objective function measures the discrepancy between the visual observations and the ones that are expected due to the generated model. The goal is to find a set of body pose parameters that minimizes the error. These pose (position and orientation) parameters define the configuration of body parts.

One drawback of these methods is the fact that initialization in the first frame of a sequence is needed since the initial estimate is based on the previous one. Another problem is the high computational cost that is associated with model-based approaches. The employed optimization method must be able to evaluate the objective function at arbitrary points in the multidimensional model parameters space, and most of the computations need to be performed online. Therefore, the resulting computational complexity is the main drawback of these methods. On the positive side, such methods are more easily extendable.

**Appearance-based methods**

Appearance-based approaches aim to establish a direct relation between image observations and pose parameters. Three main classes can be identified in this group: learning-based, example-based, and feature-based approaches. Learning-based and example-based methods do not explicitly use human body models, but implicitly model variations in pose configuration, body shape and appearance. Feature-based methods estimate pose parameters by extracting
low-level and(or) high-level features. Appearance-based methods do not need initialization step and can be used for initialization of model-based algorithms.

- **Learning-based**

These methods typically establish a mapping from image feature space to pose space using training data. The performance of these methods depends on the invariance properties of the extracted features, the number and the diversity of the body postures to be analyzed and the method utilized to derive the mapping. Due to their nature, these methods are well suited for problems such as body posture recognition where a small set of known configurations needs to be recognized. Conversely, such methods are less suited for problems that require an accurate estimation of the human body pose. Additionally, generalization for such methods is achieved only through adequate training. The main advantage of these methods is that training is performed offline and online execution is typically computationally efficient.

- **Example-based**

Instead of learning the mapping from image space to pose space, example-based methods use a database of exemplars annotated with their corresponding position and orientation data. For a given input image, a similarity search is performed and the best match is obtained from the database images. Then the corresponding pose information is retrieved as the best estimation for the actual ones. The discriminative power of these algorithms depends on the database entities. This implies in practice that body pose configurations must be densely sampled in the database entities. One of the drawbacks of these methods is the large amount of space needed to store the database. Another issue is employing efficient search algorithm to estimate the similarity between input images and database entities.

- **Feature-based**

Some methods estimate motion parameters by only extracting image features. These methods are usually developed for specific applications and cannot be generalized. Low-level features such as contours, edges, centroid of body region, and optical flow are utilized in different reported works. Low-level features are fairly robust to noise and can be extracted quickly. Another
set of methods use the location of high-level features such as fingers, fingertips, nose tip, eyes, and mouth to infer the head and hands motion parameters [61][62][63]. The major difficulty associated with feature-based methods is that body parts have to be localized prior to feature extraction.

6.3 Technical challenges

This thesis attempts to develop vision-based, marker-less methods to estimate and analyze human motion. Human motion analysis is a broad concept. Indeed, as many details as the human body can exhibit could be estimated, such as facial movement and changes in skin surface as a result of muscle tightening. In this research work, human motion analysis is limited to estimation of the global motion of head and hand, due to the fact that they are the most frequently used body parts for interacting with computers. Note that we are only interested in the estimation of the motion parameters and not the interpretation of the movements. To design and develop a robust motion estimation system, several difficulties need to be addressed:

**Motion complexity:** human motion analysis is a challenging problem due to the large variations in body motion and appearance. Although the head can be considered as a rigid object with six DOF motion, the hand is an articulated object with 27 DOF, i.e. 6 DOF for wrist (global motion) and 21 DOF for fingers (local motion). However, natural hand motion does not have 27 DOF due to the interdependences between fingers [64]. Together with the location and orientation of the hand itself, there still exist a large number of parameters to be estimated.

**Motion resolution:** since the visual resolution of the observations is limited, small changes in the body pose can go unnoticed. Body motion results in changes in a small region of the scene, and as a consequence, small movements cannot be detected.

**Rapid motion:** we humans have very fast motion capabilities. The human hand motion speed, for instance, can reach up to $5 \text{ m/s}$ for translation and $300^\circ/\text{s}$ for wrist rotation [64]. Currently, commercial cameras can support 30-60 Hz frame rates. Moreover, it is a non-trivial task for many algorithms to
achieve even a 30 Hz estimation speed. In fact, the combination of high speed body motion and low sampling rates introduces extra difficulties for motion analysis algorithms since images at consecutive frames become more uncorrelated with increasing speed of motion.

**Self-occlusions**: this problem dramatically affects the performance for hand pose estimation. Since the hand is an articulated object, its projection results in a large variety of shapes with many self-occlusions. This makes it difficult to segment different parts of the hand and extract high level features.

**Uncontrolled environments**: for widespread use, many interactive systems would be expected to operate under non-uniform backgrounds and a wide range of lighting conditions. This poses many difficulties for vision-based motion analysis algorithms.

**Processing speed**: due to the fact that the goal is to develop motion estimation systems to facilitate natural and enjoyable interaction with computers, these systems should be able to operate in real-time.
Part III

HOW
Chapter 7

Revisiting the Problem of Human Motion Analysis

In this chapter we will explain how our goal is achieved. Technical contributions of the thesis are highlighted with respect to four important aspects of the research work. We review the contributions and refer to the papers for more detailed descriptions.

7.1 New strategy; right problem, right tool

There are already many powerful tools and advanced algorithms for resolving the problem in the field. Therefore, it is not really necessary to reinvent complex techniques. What we need is innovation! That is to define the right problem (or redefine the problem), and then select right tools or integrate algorithms to form a solution.

In Paper I and Paper II we show that in a particular case the position of hand and fingertips are needed to enable natural interaction with mobile devices. In this case, we do not go for solving complicated hand gesture recognition problem. Simply, we just need to "see" the gesture as a large rotational symmetry feature. This is what we call recognition without intelligence.

This method is based on low level operators, detecting natural features without any effort to have an intelligent system. For this reason, this approach is computationally efficient and can be used in real-time applications. We propose to take advantage of rotational symmetry patterns, i.e. lines, curvatures, and circular patterns associated with the hand and fingertips that leads
Figure 7.1: Rotational symmetry patterns. From left, linear, curvature, and circular patterns with four different phases.

to the detection of the gesture. Rotational symmetries are specific curvature patterns detected from local orientation [65]. This theory was developed in 80’s by Granlund and Knutsson [66]. The main idea behind that is to use local orientation to detect complex curvatures in double-angle representation. Fig. 7.1 illustrates three orders of rotational symmetry patterns with four different phases.

7.2 New setting; active vs. passive

Considering the traditional system settings, all elements involved in the system has been set in a passive way. Input devices and sensors have been placed at a fixed position, and in the same way, users used to sit still and act in a very limited physical space passively. But we want to remove all physical constraint, allow users to move freely. This is what we aimed for; letting people to move, to be active and healthy. Therefore, we introduce a new setting, active setup, where sensors are mounted on human body instead. This concept is presented and discussed in Paper III, and also used in Paper VII. Technical question
is whether or not the problem becomes easier with this kind of active thinking and acting. The answer is Yes.

There are several issues that degrade the system performance in the passive configuration. The most essential drawback is the resolution problem. Human motion results in changes in a small region of the scene, the fact that increases the burden of detecting small movements accurately. But, we believe this problem can be easily resolved by employing the active setup. Active configuration can dramatically enhance the resolution problem. Based on the experiments in our lab, mounting the camera on the human body can enhance the resolution in the order of 10 times compared to the passive setup [67].

In order to simplify the idea, consider a simple rotation around Y-axis as it is illustrated in Fig. 7.2. This figure shows a top view of an abstract human head and a camera. Two possible configurations are presented, placing the camera at point A, in front of user (the passive setup), and mounting the camera on the head at point B (the active setup). Assume that the head rotates
with angle $\theta$. This causes a horizontal displacement of $\Delta x$ pixels of the projection of a world point, $P = (X, Y, Z)^T$, in the captured image. Assuming the perspective camera model, $\Delta x$ is given by:

$$\Delta x = \frac{f}{k} \frac{\Delta X}{Z}$$

(7.1)

where $f$ represents the focal length and $k$ the pixel size. If the camera is located at point A, then we have:

$$\Delta x_1 = \frac{f}{k} \frac{r_1 \sin \theta}{r_2 - r_1 \cos \theta}.$$  

(7.2)

Since $r \cos \theta$ is very small, $\Delta x_1$ can be written as:

$$\Delta x_1 \approx \frac{f}{k} \frac{r_1 \sin \theta}{r_2}.$$  

(7.3)

Multiplying and dividing by $\cos \theta$ yields:

$$\Delta x_1 \approx \frac{f}{k} \frac{r_1 \sin \theta \cos \theta}{r_2 \cos \theta} = \frac{f}{k} \tan \theta \frac{r_1 \cos \theta}{r_2}.$$  

(7.4)

For the active setup, where the camera is mounted on the head at point B we have:

$$\Delta x_2 = \frac{f}{k} \frac{r_2 \sin \theta}{r_2 \cos \theta} = \frac{f}{k} \tan \theta.$$  

(7.5)

$r_1 \ll r_2$, thus:

$$\frac{r_1 \cos \theta}{r_2} \ll 1 \Rightarrow \Delta x_1 \ll \Delta x_2.$$  

(7.6)

For example, if $f = 3 \, \text{mm}$, $k = 10 \, \mu\text{m}$, $r_1 = 10 \, \text{cm}$, $r_2 = 100 \, \text{cm}$, and $\theta = 45^\circ$, then the displacement for both cases will be:

$$\Delta x_1 = 3 \times 10^{-3} \times 1 \times \frac{10}{100} \times \frac{1}{\sqrt{2}} \approx 21 \, \text{pixels},$$  

(7.7)

$$\Delta x_2 = 3 \times 10^{-3} \times 1 = 300 \, \text{pixels}.$$  

(7.8)

This indicates that motion detection is much easier when mounting the camera on the head, since the active camera configuration causes changes in the entire image while the passive setup often affects a small region of the image.
7.3 New tool; 3D vs. 2D

New sensors provide a new dimension and totally change the nature of the problem. This section investigates the problem of non-linearity inherent in classical motion estimation techniques using only 2D color images, and demonstrates how new 3D sensors empower us to tackle this issue.

Here the same notations used by Horn [68] are employed to explore the nonlinearity associated with 2D RGB image-based motion estimation techniques. First, we review the equations describing the relation between the motion of a camera and the optical flow generated by the motion. We can assume either a fixed camera in a changing environment or a moving camera in a static environment. Let us assume a moving camera in a static environment. A coordinate system can be fixed with respect to the camera, with the Z-axis pointing along the optical axis. The camera motion could be separated into two components, a translation and a rotation about an axis through the origin. The translational component is denoted by \( t \) and angular velocity of the camera by \( \omega \). Let the instantaneous coordinates of a point \( P \) in the 3D environment be \((X,Y,Z)^T\). (Here \( Z > 0 \) for points in front of the imaging system.)

Let \( r \) be the column vector \((X,Y,Z)^T\), where \( T \) denotes the transpose. Then the velocity of \( P \) with respect to the \( XYZ \) coordinate system is

\[
\mathbf{V} = -\mathbf{t} - \mathbf{\omega} \times \mathbf{r}.
\]

If we define the components of \( \mathbf{V} \), \( \mathbf{t} \) and \( \mathbf{\omega} \) as

\[
\mathbf{V} = (\dot{X}, \dot{Y}, \dot{Z})^T, \quad \mathbf{t} = (U, V, W)^T, \quad \text{and} \quad \mathbf{\omega} = (A, B, C)^T
\]

we can rewrite the equation in component form as

\[
\begin{align*}
\dot{X} &= -U - BZ + CY \\
\dot{Y} &= -V - CX + AZ \\
\dot{Z} &= -W - AY + BX
\end{align*}
\]

(7.10)

where the dot denotes differentiation with respect to time.

The optical flow at each point in the image plane is the instantaneous velocity of the brightness pattern at that point [68]. Let \((x, y)\) denote the coordinate of a point in the image plane. We assume perspective projection between an object point \( P \) and the corresponding image point \( p \). Thus, the coordinates of \( p \) are

\[
x = \frac{X}{Z} \quad \text{and} \quad y = \frac{Y}{Z}.
\]
The optical flow at a point \((x, y)\), denoted by \((u, v)\) is

\[
u = \dot{x} \quad \text{and} \quad v = \dot{y}.
\]

Differentiating the equations for \(x\) and \(y\) with respect to time and using the derivatives of \(X, Y, \) and \(Z\), we obtain the following equations for the optical flow [68]:

\[
\begin{align*}
u &= \frac{\dot{X}}{Z} - \frac{X \dot{Z}}{Z^2} = (-\frac{U}{Z} - B + Cy) - x(-\frac{W}{Z} - Ay + Bx), \\
v &= \frac{\dot{Y}}{Z} - \frac{Y \dot{Z}}{Z^2} = (-\frac{V}{Z} - Cx + A) - y(-\frac{W}{Z} - Ay + Bx).
\end{align*}
\]

The resultant equations for the optical flow are inversely proportional to the distance of \(P\) to the camera \((Z)\). Unlike the motion parameters \((A, B, C, U, V, W)\) which are global and point independent, \(Z\) is pointwise and varies at each point. Therefore, \(Z\) should be eliminated from the optical flow equations.

After removing \(Z\) from the equations, we eventually obtain the following equation at each point:

\[
\begin{align*}
x(U C - W v + AW) + y(V C + W u + BW) + x y(B U + V A) - y^2(C W + A U) \\
-x^2(V B + C W) - V(B + u) + U(v - A) &= 0.
\end{align*}
\]

Here the problem arises, since the final equation is nonlinear and pointwise in terms of \(u\) and \(v\). However, this issue can be simply resolved by utilizing depth information acquired by new 3D sensors. In Paper IV, Paper V, and Paper VI we will show how new 3D sensors enable us to develop 3D linear motion estimation method.

7.4 New perspective; search-based motion estimation

Due to the increasing capability of computers in storing and processing extremely large databases, we introduce a novel approach, and propose to shift the complexity from computer vision and pattern recognition algorithms to large-scale image retrieval methods. The idea is to redefine the motion estimation problem as the image search problem. The new search-based method
is based on storing an extremely large database of body gesture images and retrieving the best match from the database. The most important point is that the database should include body gesture images annotated with their corresponding position and orientation information.

The block diagram of the search-based method is illustrated in Fig. 7.3. It includes four main components: annotated and indexed gesture database, image query processing unit, search engine, and interface level. The query image will be processed and its visual features will be extracted for similarity analysis. The search engine retrieves the best image based on the similarity of the query image with the database entries. Eventually, the best match and its position and orientation information will be used to facilitate different applications. This approach might seem similar to example-based methods; however, there is a systematic difference between them. The main idea is to form a large database, hundred thousands of images, and store them on rather too small space. This is one of the challenging issues in example-based methods. In addition, in the proposed method, searching the large-scale database can be performed more efficiently in real-time.

This method is filed as a U.S. patent application. It has successfully passed the novelty analysis step, and is currently undergoing necessary processing. Due to the patent application restrictions, the detailed description of this method cannot be included in this thesis.
Chapter 8

Comparative Evaluation of Developed Systems

Technical contributions of this research work have resulted in developing the following systems:

- 3D interaction with mobile devices (Paper I and Paper II)
- Active motion estimation system (Paper III)
- Direct motion estimation system (Paper IV)
- Depth-based gestural interaction system (Paper VI)

In this section, the developed systems will be quantitatively and qualitatively compared to state-of-the-art and the results will be presented.

8.1 3D interaction with mobile devices

This system provides natural interaction with mobile devices. The system is developed for finger tracking in Paper I, and is improved to extract 3D hand motion in Paper II. Extensive experiments have been carried out to evaluate the performance of the system, both on desktop and mobile platforms. The experimental results are discussed in Paper I and Paper II. Here we try to provide comparative evaluation.

Due to the high computational cost, most available marker-less systems cannot perform real-time gesture analysis on mobile devices [70] [71] [72].
Table 8.1: Comparison between the proposed mobile interaction system and other vision-based works (NA: not available).

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Approach</th>
<th>Real-time</th>
<th>Detection rate/error</th>
<th>Stationary /mobile</th>
<th>Tracking /3D motion</th>
</tr>
</thead>
<tbody>
<tr>
<td>[70]</td>
<td>Model-based</td>
<td>No</td>
<td>NA / NA</td>
<td>Yes / No</td>
<td>Yes / Yes</td>
</tr>
<tr>
<td>[71]</td>
<td>Model-based</td>
<td>No</td>
<td>NA / NA</td>
<td>Yes / No</td>
<td>Yes / Yes</td>
</tr>
<tr>
<td>[72]</td>
<td>Learning-based</td>
<td>Yes</td>
<td>NA / NA</td>
<td>Yes / No</td>
<td>Yes / No</td>
</tr>
<tr>
<td>[74]</td>
<td>Feature-based</td>
<td>Yes</td>
<td>≈ 98% / NA</td>
<td>Yes / No</td>
<td>Yes / No</td>
</tr>
<tr>
<td>[73]</td>
<td>Feature-based</td>
<td>Yes</td>
<td>≈ 84% / NA</td>
<td>Yes / No</td>
<td>Yes / No</td>
</tr>
<tr>
<td>[75]</td>
<td>Feature-based</td>
<td>Yes</td>
<td>NA / NA</td>
<td>Yes / Yes</td>
<td>Yes / No</td>
</tr>
<tr>
<td>Our system</td>
<td>Feature-based</td>
<td>Yes</td>
<td>≈ 100% / 6.6 pixels</td>
<td>Yes / Yes</td>
<td>Yes / Yes</td>
</tr>
</tbody>
</table>

[73]. However, we can evaluate our technical contribution from different aspects of detection, tracking, and 3D motion analysis. Our algorithm always detects the gesture with the flexibility of the ±45 degrees rotation around x, y, and z axes within the interaction space. This detection rate is quite promising in comparison with the 84.4% and 98.42%, proposed by [73] and [74] (both algorithms only work on stationary systems). Moreover, the mean error of the localized gesture in our algorithm is 6.59 pixels which is substantially improved (reduced to half), from the proposed system by [72]. In addition, our system provides 3D motion analysis by retrieving the rotation parameters between the consecutive image frames. Since the relative distance between the user’s gesture and the device’s camera is rather small, this configuration is almost the same as active configuration, proposed in Paper III. This accurate system can guarantee the effective 3D manipulation in virtual and augmented reality environments. In Table 8.1, characteristics of the proposed system and other vision-based related works are compared from different aspects.
Table 8.2: Comparative evaluation of the proposed active motion estimation approach (NA: not available, MAE: mean absolute error).

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Approach</th>
<th>Features</th>
<th>Camera</th>
<th>MAE</th>
<th>Real-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>[76]</td>
<td>Active configuration</td>
<td>SIFT</td>
<td>Multi-camera</td>
<td>3.0°</td>
<td>No</td>
</tr>
<tr>
<td>Our system</td>
<td>Active configuration</td>
<td>SIFT</td>
<td>Single</td>
<td>0.8°</td>
<td>Yes</td>
</tr>
</tbody>
</table>

8.2 Active motion estimation

Active configuration for human motion analysis is introduced in Paper III. By introducing a new setting, we propose to mount optical sensors onto the human body to measure and analyze body movements. Our system is camera-based and relies on the rich data in a detailed view of the environment. This system recovers motion parameters by extracting stable features in the scene and finding corresponding points in the consecutive frames.

While this approach has previously been attempted in visual odometry for localization in GPS-blind environments [77] [78] [79] and Simultaneous Localization and Mapping (SLAM) for estimating the motion of moving platforms [80] [81] [82], we propose to take the advantage of active configuration for human motion estimation.

A closely related work is that of Shiratori et al. [76]. In both works active set-up is used to estimate human motion, and color camera is regarded as the only employed sensor. However, Shiratori utilizes a set of cameras attached onto the body to estimate the pose parameters. Thus, instead of recovering the independent ego-motion of individual cameras, they try to reconstruct the 3D motion of a set of cameras related by an underlying articulated structure.

To quantify the angular accuracy of the system, we have provided an electronic measuring device which is presented in Paper III. Outputs of this measuring device are considered as the Ground Truth to evaluate the system performance. The device consists of a protractor, a servo motor with an indicator, and a control board connected to a power supply. A normal webcam is also fixed on the servo motor, so its rotation is synchronized with the servo. The servo can move in two different directions with specified speed, and its actual
rotation value (the Ground Truth) is indicated on the protractor. Whenever the servo motor rotates, the camera will be rotated and the video frames will be fed to the system to estimate the camera rotation parameters. These values are used to measure the accuracy of the system. Three different setups are used to analyze the system performance in rotations around X, Y, and Z axes. The servo motor can be operated through Lynx SSC-32 Terminal [83]. This servo controller has a range of $180^\circ$ with a high resolution of $0.09^\circ$, and can be used for accurate positioning. The performance of the system is compared to that of [76] and is presented in Table 8.2.

### 8.3 Direct motion estimation

One of the objectives of this thesis is to explore the opportunity provided by new 3D sensors. Range sensors provide depth information and introduce new possibilities for developing motion estimation techniques. By employing depth images, we introduce a direct method for 3D rigid motion estimation in Paper IV.

Due to the availability of Ground-Truth data, the performance of the direct motion estimation system can be evaluated in comparison with other methods. Two different approaches, model-based [84] and learning-based [85] are used for performing comparative evaluation. In [84], authors have used a 3D head model to estimate the pose parameters. The head pose estimation problem has been formulated as an optimization problem that has been solved based on Particle Swarm Optimization. Instead of using a head model, [85] utilizes random regression forests to establish a mapping between depth features and head pose parameters.

Biwi database [86] has been used to evaluate the methods. The database contains over 15K depth images (640 x 480 pixels) of 20 people. For each depth image, the Ground-Truth head motion parameters are provided. Table 8.3 illustrates mean and standard deviation of the errors for the 3D head localization task and the individual rotation angles with required processing time. Although the model-based approach provides higher accuracy, but it is computationally costly and is not favorable for real-time applications. On the other hand, the learning-based method performs beyond real time, though offering low accuracy. In contrast to these systems, our approach provides acceptable accuracy with low processing time that makes this method suitable for real-
Table 8.3: Properties of the model-based and learning-based methods for the head pose estimation problem are compared with the proposed approach of this thesis.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Approach</th>
<th>Localization error (mm)</th>
<th>Pitch error</th>
<th>Yaw error</th>
<th>Roll error</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>[84]</td>
<td>Model-based</td>
<td>2.8 ± 1.8</td>
<td>1.3 ± 1.1°</td>
<td>1.1 ± 1.0°</td>
<td>1.7 ± 1.7°</td>
<td>100 ms</td>
</tr>
<tr>
<td>[85]</td>
<td>Learning-based</td>
<td>14.6 ± 22.3</td>
<td>8.5 ± 9.9°</td>
<td>8.9 ± 13.0°</td>
<td>7.9 ± 8.3°</td>
<td>25 ms</td>
</tr>
<tr>
<td>Our method</td>
<td>Feature-based</td>
<td>10.2 ± 17.6</td>
<td>6.4 ± 7.9°</td>
<td>6.3 ± 8.1°</td>
<td>6.5 ± 8.8°</td>
<td>40 ms</td>
</tr>
</tbody>
</table>

time applications.

8.4 Depth-based gestural interaction

In Paper VI, a novel approach for performing gestural interaction is presented. By using the direct motion estimation technique, we developed a system that is able to extract the global hand motion parameters. The hand pose parameters can be employed to provide natural interaction with computers. Several experiments are conducted to demonstrate the system accuracy and capacity to accommodate different interaction procedures, which are discussed in more detail in Paper VI.

In this part, the proposed depth-based gestural interaction system is compared to other depth-based approaches with respect to different aspects. Three model-based [41] [42] [87] and two example-based [88] [89] systems are selected for the evaluation. The results of the comparative analysis are presented in Table 8.4. Model-based approaches estimate global hand motion and full joint angles. Note that in [42] the pose parameters are estimated for two hands.

Model-based methods are computationally costly, and cannot be used in real-time applications. However they provide continuous solutions and can be generalized. Despite the fact that example-based methods are well suited for real-time applications, they only estimate hand position or orientation, and are not extendible. In comparison with these systems, our system is able to estimate global hand motion parameters continuously and fast, which is desirable for real-time applications. Like model-based systems, the proposed system can be generalized to other manipulative gestures.
Table 8.4: Comparison between the proposed depth-based gestural interaction system and other approaches.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Approach</th>
<th>Estimated parameters</th>
<th>Continuous estimation</th>
<th>Complexity</th>
<th>Extendible</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>[41]</td>
<td>Model-based</td>
<td>Global hand motion + joint angles</td>
<td>Yes</td>
<td>High</td>
<td>Yes</td>
<td>40 s</td>
</tr>
<tr>
<td>[42]</td>
<td>Model-based</td>
<td>Global hand motion + joint angles</td>
<td>Yes</td>
<td>High</td>
<td>Yes</td>
<td>250 ms</td>
</tr>
<tr>
<td>[87]</td>
<td>Model-based</td>
<td>Global hand motion + joint angles</td>
<td>Yes</td>
<td>High</td>
<td>Yes</td>
<td>66 ms</td>
</tr>
<tr>
<td>[88]</td>
<td>Example-based</td>
<td>Hand orientation</td>
<td>No</td>
<td>Low</td>
<td>No</td>
<td>34 ms</td>
</tr>
<tr>
<td>[89]</td>
<td>Example-based</td>
<td>Hand position</td>
<td>No</td>
<td>Low</td>
<td>No</td>
<td>34 ms</td>
</tr>
<tr>
<td></td>
<td>Our system</td>
<td>Feature-based</td>
<td>Yes</td>
<td>Low</td>
<td>Yes</td>
<td>36 ms</td>
</tr>
</tbody>
</table>
Chapter 9

Summary of the Selected Articles

This thesis is based on the contributions of seven papers which are referred to in the text and presented in the last part of the thesis. **Paper I** results from an equal collaboration between the first author and Farid Abedan Kondori. In the other papers Farid Abedan Kondori is the main author and has major contributions in developing theories, implementations, experiments, and writing. Associate Professor Li Liu and Professor Haibo Li have inspired the discussions and have supervised Farid Abedan Kondori during the PhD study. Shahrouz Yousefi, PhD candidate, and Jean-Paul Kouma, licentiate of engineering in computer vision, have participated in the discussions and have assisted Farid Abedan Kondori in some experiments.

**Paper I**
Shahrouz Yousefi, **Farid Abedan Kondori**, Haibo Li

*Camera-based gesture tracking for 3D interaction behind mobile devices.*

In this paper we demonstrate how position of fingertips can be utilized to provide natural interaction with mobile devices. This paper suggests the use of first order rotational symmetry patterns, curvatures, to detect fingertips. This method is based on low level operators, detecting natural features associated with fingertips without any effort to have an intelligent system, i.e. recognition without intelligence. The proposed approach is computationally efficient and
can be used in real-time applications.

**Paper II**

Farid Abedan Kondori, Shahrouz Yousefi, Haibo Li  
*Real 3D interaction behind mobile phones for augmented environments.*  
In Proceeding of the IEEE International Conference on Multimedia and Expo (ICME2011), Barcelona, Spain, July 2011.  
This paper suggests employing 3D hand gesture motion to enable a new way to interact with mobile phones in 3D space behind the device. Hand gesture recognition is achieved by detecting second order rotational symmetry patterns, i.e. circular patterns. Relative 3D motion of hand gesture between consecutive frames is estimated by means of extracting stable features in the scene. We demonstrate how 3D hand motion recovery provides a novel way to manipulate objects in augmented reality applications.

**Paper III**

Farid Abedan Kondori, Li Liu  
*3D active human motion estimation for biomedical applications.*  
In Proceeding of the World Congress on Medical Physics and Biomedical Engineering (WC2012), Beijing, China, May 2012.  
In this paper, a vision-based active human motion estimation system is developed to diagnose and treat movement disorders. One of the important factors in healthcare is to observe and monitor patients in their daily life activities. However current motion analysis systems are confined to lab environments and are spatially limited. This hinders the possibility of observing patients while moving naturally and freely, the matter that can affect the quality of the diagnosis. By introducing a new setting, we propose to mount cameras on human body to measure and analyze body movements. The system extracts stable features in the scene and finds corresponding points in the consecutive frames. Afterwards it recovers 3D motion parameters.

**Paper IV**

Farid Abedan Kondori, Shahrouz Yousefi, Haibo Li  
*Direct 3D head pose estimation from Kinect-type sensors.*  
In this paper, it is shown how to take the advantage of new 3D sensors to introduce a direct method for estimating 3D motion parameters. Based on the
range images, we derive a new version of optical flow constraint equation to directly estimate 3D motion parameters without any need of imposing other constraints. In this work, the range images are treated as classical intensity images to obtain the new constraint equation. The new optical flow constraint equation is employed to recover 3D motion of a moving head from the sequences of range images. Furthermore, a closed-loop feed-back system is proposed to handle the case when optical flow is large.

**Paper V**

**Farid Abedan Kondori, Shahrouz Yousefi, Li Liu, Haibo Li**

*Head operated electric wheelchair.*

In Proceeding of the IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI2014), San Diego, California, USA, April 2014.

In order to steer and control an electric wheelchair, this paper proposes to utilize user head orientation. Most of the electric wheelchairs available in the market are joystick-driven, and therefore assume that users are able to use their fine motor skill to drive the wheelchair. Unfortunately, this is not valid for many patients like those suffering from quadriplegia. These patients are usually unable to control an electric wheelchair using conventional methods, such as joystick or chin stick. This paper introduces a novel assistive technology to help these patients. By acquiring depth information from Kinect, user head orientation is estimated and mapped onto control commands to steer the wheelchair.

**Paper VI**

**Farid Abedan Kondori, Shahrouz Yousefi, Jean-Paul Kouma, Li Liu, Haibo Li**

*Direct hand pose estimation for immersive gestural interaction.*

Submitted to Pattern Recognition Letters, February 2014.

This paper presents a novel approach for performing intuitive gestural interaction. The main issue, i.e. dynamic hand gesture recognition, is addressed as a combination of two tasks: gesture recognition and gesture pose estimation. The former initiates the gestural interaction by recognizing a gesture as belonging to a predefined set of gestures, while the latter provides required information for recognition of the gesture over the course of interaction. To ease dynamic gesture recognition, this paper proposes a direct method for hand pose recovery. We demonstrate the system performance in 3D object manipulation.
on two different setups; desktop computing and mobile platform. This reveals the system capability to accommodate different interaction procedures. Additionally, usability test is conducted to evaluate learnability, user experience and interaction quality in 3D gestural interaction in comparison to 2D touch-screen interaction.

**Paper VII**

**Farid Abedan Kondori, Li Liu, Haibo Li**

*Telelife; a new concept for future rehabilitation systems?*

Submitted to the IEEE Communications Magazine, April 2014.

In this paper we introduce telelife, a new concept for future rehabilitation systems. Recently, telerehabilitation systems for home-based therapy have altered healthcare systems. The main focus of these systems is on helping patients to recover physical health. Although telerehabilitation provides great opportunities, there are two major issues that affect effectiveness of telerehabilitation: relegation of patient at home, and loss of direct supervision of therapist. To complement telerehabilitation, this paper proposes telelife. Telelife can enhance patients’ physical, mental, and emotional health. Moreover, it improves patients’ social skills and also enables direct supervision of therapists. In this work we introduce telelife to enhance telerehabilitation, and investigate technical challenges and possible solutions to achieve it.

### 9.1 List of all publications

All the contributions that have been published during the PhD study are listed here.

**Journal articles**

- **Farid Abedan Kondori, Li Liu, Haibo Li.** *Telelife; a new concept for future rehabilitation systems?* Submitted to the IEEE Communications Magazine, 2014.

- **Farid Abedan Kondori, Shahrour Yousefi, Jean-Paul Kouma, Li Liu, Haibo Li.** *Direct hand pose estimation for immersive gestural interaction.* Submitted to Pattern Recognition Letters, 2014.


**Conference papers**


• **Farid Abedan Kondori**, Shahrouz Yousefi, Li Liu, Haibo Li. *Head operated electric wheelchair*. In Proceeding of the IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI2014), April 6-8, 2014, San Diego, California, USA.


- Shahrouz Yousefi, **Farid Abedan Kondori**, Haibo Li. *3D gestural interaction for stereoscopic visualization on mobile devices*. In Proceeding of the 14th International Conference on Computer Analysis of Images and Patterns (CAIP’11), Seville, Spain, August 29-31, 2011, pp.555-562.


- Shahrouz Yousefi, **Farid Abedan Kondori**, Haibo Li. *Stereoscopic visualization of monocular images in photo collections*. In Proceeding of the IEEE International Conference on Wireless Communications and

• Shahrouz Yousefi, **Farid Abedan Kondori**, Haibo Li. *Tracking fingers in 3D space for mobile interaction*. In Proceeding of The 20th International Conference on Pattern Recognition (ICPR), The Second International Workshop on Mobile Multimedia Processing (WMMP), Istanbul, Turkey, August 2010, pp.72-79.

**Licentiate thesis**

• **Farid Abedan Kondori.** *Human motion analysis for creating immersive experiences*. Umeå University, Umeå, Sweden, April 2012.

**Patents**

Chapter 10

Contributions, Conclusions, and Future Directions

10.1 Contributions

Contributions of this thesis can be divided into two parts: technical contributions towards body motion analysis, and implementation of the developed methods for providing bodily interaction in various application scenarios.

10.1.1 Body motion analysis

Technical contributions of this thesis are focused on body gesture detection, tracking, and analysis. These contributions can be listed as:

- Hand gesture recognition and tracking based on low-level operations.

- Relative motion estimation between consecutive frames by extracting and tracking stable landmarks in the environment.

- Adopting active configuration for human motion analysis.

- Taking advantage of 3D sensors to develop a direct and linear motion estimation method.

- Introducing a novel approach for body gesture analysis based on large-scale search framework.
10.1.2 Bodily interaction

The proposed motion analysis systems can be utilized to facilitate a healthy, intuitive, and pleasurable interaction in different applications. In these applications, gesture motion is captured, and detailed analysis of the motion is used to provide different functionalities. The application scenarios are summarized in the following:

- 3D gestural interaction with mobile devices.
- Bare-hand interaction with computers for 3D object manipulation.
- Gestural interaction for browsing medical images in the operating rooms.
- Assistive technology for controlling electric wheelchairs.
- Analysis of human daily life activities for diagnosing and treating movement disorders.

10.2 Conclusions

This thesis aimed to propose bodily interaction methods that empower human to have a healthy, natural, and enjoyable interaction with computers. In order to achieve our aim, we have developed computer vision-based, marker-less systems that use low-level features to detect, track, and analyze body gestures.

- We presented a novel approach for recovering six degrees of freedom of hand motion. This approach is computationally efficient and relies on low level operations, detecting natural features with the minimum level of intelligence. Due to the low complexity, this method can be used in real-time applications to estimate the hand motion. However, rapid hand motions affect the system performance.

- This thesis proposed an active configuration to capture and analyze body motion. Mounting the camera on the body, we have built a convenient, low cost motion capture system that can be used even in daily-life activities. This system provides high resolution and can be used with high accuracy in real-time applications. Detecting sufficient number of interest points in the scene is the only requirement for the proposed method.
• Moreover, we derived a new version of optical flow constraint equation for range images. This new equation not only enables a direct recovery of 3D motion, but also offers an opportunity of re-using the techniques, tools, and experiences developed and accumulated in the topic of optical flow computation in the computer vision fields. Simply by treating range images as conventional intensity images, we can apply the proved techniques and tools used in the optical flow computation for recovering 3D motion parameters. Obviously, more experiments need to be done to verify if the new equation could act as a fundamental equation in the coming era of range data just as the role of the normal optical flow equation in motion estimation from 2D intensity images.

• In addition, the main challenge for providing bare-hand gestural interaction, i.e. dynamic gesture recognition, has been discussed in this thesis. This issue is regarded as a combination of two tasks: gesture recognition and gesture motion estimation. We proposed to incorporate a fast and robust hand pose estimation to ease the problem.

• We also demonstrated that advanced algorithms from other research areas could be employed to tackle the problem of human motion analysis. In order to handle the diversity and flexibility of the body movements, we decided to shift the complexity from classical motion estimation problem to large-scale image retrieval system. Due to the availability of the large-scale image databases, the new problem is to find the best match for the query among the whole database. In fact, for a query image or video, representing a unique body gesture with specific position and orientation, the challenging task is to retrieve the most similar image from the database that represent the same gesture with maximum similarity in position and orientation. In this approach, annotating the database images with their actual pose parameters is crucially important.
10.3 Future directions

Our vision is to improve human well-being by means of a natural and enjoyable interaction. Vision-driven design\(^1\) can result in quantum leaps in the way we interact with computers. It complements needs-driven and technology-driven design by looking beyond today limits. Focusing on user needs and centralizing what user wants from technology has been essential for designing interaction methods. However, new vision extents this view and concerns how human values can be supported through new technology.

The motivation for this work was introducing bodily interaction methods that will support human well-being. The main reason why we focus on this vision-driven approach is its life span. Technologies can become outdated in a few years, and user needs can change quickly and applications become obsolete within a decade. However, vision-driven approaches can last for a long time. Improving human physical, mental, emotional, and social well-being will be central for future technologies. Bodily interaction with computers can enhance human well-being. As a result, utilizing human body in different situations will be a common approach in design and development of future interaction methods.

Although we showed the high potential of computer vision-based systems, using other sensing modalities can facilitate bodily interaction systems. Indeed, bodily interaction methods can benefit from the fusion of vision- and non-vision-based systems.

---

\(^1\)Note the difference between vision-driven design and vision-based motion estimation. Vision-driven design refers to our vision, i.e. improving human well-being.
BIBLIOGRAPHY
Bibliography


[16] http://structure.io/#/.


[85] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, “Real time head pose estimation from consumer depth cameras,” in 33rd Annual Symposium of the German Association for Pattern Recognition (DAGM’11), September 2011.

