Human Motion Analysis

for Creating Immersive Experience

FARID ABE DAN KONDORI

Digital Media Lab
Department of Applied Physics and Electronics
Umeå University
Umeå, Sweden 2012
Human Motion Analysis
for Creating Immersive Experience

Farid Abedan Kondori
Department of Applied Physics and Electronics, Umeå University

ABSTRACT
From an early age, people display the ability to quickly and effortlessly interpret the orientation and movement of human body parts, thereby allowing one to infer the intentions of others who are nearby and to comprehend an important nonverbal form of communication. The ease with which one accomplishes this task belies the difficulty of a problem that has challenged computational systems for decades, human motion analysis.

Technological developments over years have resulted into many systems for measuring body segment positions and angles between segments. In these systems human body is typically considered as a system of rigid links connected by joints. The motion is estimated by the use of measurements from mechanical, optical, magnetic, or inertial trackers. Among all kinds of sensors, optical sensing encompasses a large and varying collection of technologies.

In a computer vision context, human motion analysis is a topic that studies methods and applications in which two or more consecutive images from an image sequences, e.g. captured by a video camera, are processed to produce information based on the apparent human body motion in the images.

Many different disciplines employ motion analysis systems to capture movement and posture of human body for applications such as medical diagnostics, virtual reality, human-computer interaction etc.

This thesis gives an insight into the state of the art human motion analysis systems, and provides new methods for capturing human motion.

Keywords: Human Motion Analysis, Active Motion Capture, Passive Motion estimation, 3D Head Pose Estimation, Hand Gesture Recognition, Hand Gesture Motion Estimation, Human Computer Interaction, Immersive Interaction.
Preface

This thesis is based on the following publications:

I. Farid Abedan Kondori, Li Liu, “3D Active Human Motion Estimation for Biomedical Applications,” accepted in World Congress on Medical Physics and Biomedical Engineering (WC2012), Beijing, China, 26-31 May 2012.


Other publications are listed here:


- Farid Abedan Kondori, Shahrouz Yousefi, “Smart Baggage in Aviation,” in IEEE International Conference on Internet of Things and International


Under-review journal articles


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 supervisor Prof. Haibo Li for not only the opportunities provided and all the hard work, but also for the positive environment he creates in our research group. I am truly indebted and thankful for the valuable guidance from him in the research world and for all the support and encouragement.

I am grateful to my advisors, Dr. Adi Anani and Dr. Li Liu for inspiring discussions, practical suggestions, and providing the financial support.

I would also like to thank my best friend, and my best colleague, Shahrouz Yousefi with whom I have spent a great deal of time since 2001. Without his contributions I would not have been able to pursue my research.

Special gratitude goes to my friends and colleagues, Zeynab Kolahi, Jean-Paul Kouma, Shafiq Ur Rehman, Alaa Halawani, and Ulrik Söderström for their inspiration and support.

Thanks to all the staff at the Department of Applied Physics and Electronics (TFE) for creating an enjoyable and interesting working environment. I express my warm thanks to Annemaj Nilsson and Mona-Lisa Gunnarsson who have always helped me.

Special thanks also to Annika Bindler, who gave a great deal of assistance for correcting the writing errors in the final revision of this thesis.

Finally but 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.

Thank you all.

Farid Abedan Kondori
Umeå, April 2012
ACKNOWLEDGMENTS
Contents

Abstract i
Preface iii
Acknowledgments v

I INTRODUCTION 1

1 Introduction 3
   1.1 Motivation 3
   1.2 State of the art 4
   1.3 New areas, new possibilities 5
   1.4 Research goal 8
   1.5 Potential impact 8
   1.6 Thesis outline 9

II DEVELOPING MOTION CAPTURE TECHNIQUES 11

2 Head Pose Estimation 13
   2.1 Introduction 13
   2.2 Related work 14
   2.3 Active & Passive motion capture 15
   2.4 Active head motion estimation 16
      2.4.1 Head pose estimation 17
   2.5 Passive head motion estimation 19
      2.5.1 3D linear & 2D nonlinear methods 19
      2.5.2 Kinect 21
      2.5.3 System description 21
      2.5.4 Multiple head detection and tracking 22
## CONTENTS

### 3 Hand Motion Estimation
- 3.1 Introduction .................................................. 27
- 3.2 Previous implementation ........................................ 28
- 3.3 Vision-based gesture detection & tracking ..................... 30
  - 3.3.1 Gesture detection ........................................... 30
  - 3.4 Gesture motion estimation ..................................... 31

## III EVALUATING MOTION CAPTURE SYSTEMS

### 4 Evaluation of Head Motion Estimation Systems
- 4.1 Introduction .................................................. 35
- 4.2 Design criteria .................................................. 35
- 4.3 System evaluation .............................................. 36
  - 4.3.1 Passive head motion estimation ............................. 36
  - 4.3.2 Active head motion estimation ............................ 38
- 4.4 Conclusion .................................................... 40

### 5 Evaluation of Hand Gesture Recognition System
- 5.1 Introduction .................................................. 43
- 5.2 Design criteria .................................................. 43
- 5.3 System Evaluation .............................................. 44

## IV APPLICATION DOMAINS

### 6 Applications
- 6.1 Application areas .............................................. 49

## V CONTRIBUTIONS, DISCUSSIONS, AND FUTURE DIRECTIONS

### 7 Outline and Summary of Contributions
- 7.1 Summary of contributed papers ............................... 55
- 7.2 Summary of contributions ..................................... 57

### 8 Concluding Remarks & Future Directions
- 8.1 Conclusions .................................................... 59
- 8.2 Future directions ............................................... 60
List of Figures

1.1 Marker-based human motion analysis. Left: Schematic of the set-up with six cameras. Right: Subject equipped with reflective markers that are detected and tracked by cameras to estimate the motion [1]. .................................................. 5

1.2 Emergence of new visual sensors, advance computer vision algorithms, and applications that require immersive interactions show us the future directions to develop new human motion analysis techniques. .................................................. 6

1.3 Moving from 2D touch screen interaction space towards 3D space behind smartphones, new natural user interfaces for mobile applications are designed. Users can manipulate the virtual object on the mobile screen using their hand gestures [2]. ................. 8

2.1 Top view of a head and a fixed camera. The head turns with angle $\theta$ causing a change in the resulted image. The amount of change depends on the camera location (A or B) [3]. .......................... 15

2.2 Active motion tracking system. ................................. 17

2.3 Top, two consecutive frames. Bottom, point correspondences between them [3]. .............................. 18

2.4 Microsoft Kinect: (A) laser projector, (B) RGB camera, (C), monochrome CMOS camera [4]. ................................. 20

2.5 Passive head pose estimation system [4]. .................. 21

2.6 Head detection scheme [4]. ................................. 23

2.7 Head localization [4]. ................................. 24

3.1 Taxonomy of hand gestures. ................................. 27

3.2 3D vision-based human mobile interaction system [2]. .... 29

3.3 Rotational Symmetries. Linear, curvature, and circular patterns [2]. 30

3.4 (a) User hand gesture, (b) Localized gesture [2]. ............... 31
3.5 Feature matching in two consecutive frames where 54 point cor-
respondences are detected [2]. . . . . . . . . . . . . . . . . . . . . 32

4.1 Experimental results. The 3D head motion parameters are esti-
mated to update the position and orientation of the 3D model. First row is the initial position. Next three rows show the rotation around X, Y, and Z axes respectively. The last three rows illustrate the translation in X, Y, and Z [4]. . . . . . . . . . . . . . . 37

4.2 Electronic measuring device. a) The setup for Z-axis, b) for X-
axis, and c) for Y-axis [3]. . . . . . . . . . . . . . . . . . . . . . 38

4.3 Active motion tracking demo. As the user turns his head, the mo-
tion parameters are estimated and used to change the 3D model on the computer screen [3]. . . . . . . . . . . . . . . . . . . . . . 39

5.1 (a) System performance in gesture tracking. (b) Error of the tracking in a sequence of images [5]. . . . . . . . . . . . . . . 44

5.2 Relative rotation and translation between image sequences. (a) The teapot should rotate in the direction specified by the arrow while the user gesture rotates in the same direction. (b) Consecutive frames (top), feature matching between two images (middle), and applying user gesture motion to the teapot (bottom) [2]. . . 45
Part I

INTRODUCTION
Chapter 1

Introduction

1.1 Motivation

The study of human motion dates back to 1870s when Muybridge [6] 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.

Human motion analysis plays an essential role in human computer interaction (HCI), for bridging the information gap between humans and computers. Currently, due to the large influx of computers in our daily lives, HCI has become crucially important. Conventionally, keyboard and mouse have played the main role in HCI. However, with the rapid progress of computing, communication, and display technologies, such interfaces may become a bottleneck in applications that rely on heavy interaction of users with machines because of the unnaturalness of the interaction. Thus, researchers have recently attempted to eliminate this HCI bottleneck by developing more natural ways of interaction. With this motivation, human motion estimation and human gesture recognition have been topics of research for decades.

Additionally, tracking human body parts and recovering the underlying 3D human body structure are critically valuable for medical diagnostics systems, entertainment industry, and analysis of athletic performance. The capability to automatically observe human activities in security-sensitive areas such as airports, borders, and banks is of great interest to the security services as well.
CHAPTER 1. INTRODUCTION

1.2 State of the art

Most of the existing human motion tracking and analysis systems can be classified into two categories: position sensing systems and vision-based tracking systems.

- **Position sensing systems**

  In the position sensing paradigm, a set of sensors is mounted on the user 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 [7] [8]. It is sensitive to vibrational artifacts [9]. Another shortcoming of the accelerometers is the lack of information about the rotation around the global Z-axis [10]. 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 [11]. Although, the major disadvantage of the 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.

  Magnetic sensors can be utilized for human motion estimation as well. The use of the magnetic sensors is reported in several works [12] [13]. A magnetic sensor, or magnetometer, is a device that is used to measure the strength and direction of a magnetic field. The performance of the magnetic sensors is affected by the availability of ferromagnetic materials in the surrounding environment [14].

- **Vision-based systems**

  Vision-based motion capture systems rely on a camera as an optical sensor. 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. One example of such a system is illustrated in Fig. 1.1. These systems are accurate and have been used successfully in biomedical applications [15] [16] [17]. Although, many difficulties are associated with such a configuration. For instance, scale changes (distance of the user to the 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
Figure 1.1: Marker-based human motion analysis. Left: Schematic of the set-up with six cameras. Right: Subject equipped with reflective markers that are detected and tracked by cameras to estimate the motion [1].

Marker-less systems rely only on cameras and try to employ computer vision techniques to estimate the motion. The use of cheap cameras is possible in such systems [18] [19]. However, the markers removal comes with the price of complicating the estimation process of 3D non-rigid human motion. Developing marker-less motion capture systems is still an on-going research topic in computer vision, and only partial success in real situations has been achieved [20]. Several problems such as cluttered scene, human occlusion, scale variation and illumination can degrade the system performance. Nevertheless, the most essential drawback reported in these systems is the resolution problem. Since human body motion results in changes in a small region of the scene, small movements cannot be detected.

1.3 New areas, new possibilities

In addition to the limitations of the previous implementations, it needs taken into account that with the recent progress in technology and computer science, there are new factors that offer new possibilities to research communities for developing
new human motion analysis theories. Three main areas that provide us with new opportunities to develop new motion analysis techniques are new visual sensors, fast and robust computer vision algorithms, and emergence of new applications (Fig. 1.2).

- **Sensor-driven methodologies**

New visual sensors have opened a new angle for researchers recently. New high resolution cameras are becoming smaller, and therefore, can be conveniently mounted onto a human body to analyze human activities. They are also capable of recording human daily life activities for future analysis. Hence, one can ask the question: why not use small, low-cost cameras as optical trackers to estimate human motion?

Another type of inexpensive visual sensors that can be employed to analyze human movements is new depth sensors. Kinect is one of those new depth sensors [21]. It interprets 3D scene information from a continuously-projected infrared structured light. Kinect provides a robust solution to infer 3D information of the scene and has a big potential to be used in human motion analysis systems. Using the knowledge about 3D structure of the scene, it is possible to directly estimate human motion. Chapter 2 presents an in-depth discussion concerning this area.

- **Algorithm-driven methodologies**

During the last decade, massive contributions and publications in the computer vision field have resulted into robust algorithms that have high potential to be utilized in human motion analysis systems. In recent years, scale-invariant feature
transform (SIFT) [22] has become a strong tool for researchers to overcome the traditional limitations in many areas in computer vision fields, such as object detection, classification, and motion estimation. However, since SIFT is computationally expensive, it was impossible to apply it to real-time applications. Fortunately, today, with integrating graphical processing units (GPUs) in computers and mobile phones, it is possible to efficiently implement this robust computer vision algorithm into computers and mobile applications.

- Application-driven methodologies

New possibilities for creating new motion analysis systems have become even more apparent with the emergence of smart phones, 3D TVs, and intelligent environments (IEs). Almost all new smart phones are equipped with powerful processors, as well as high resolution cameras. Thus, they provide the opportunity to employ computer vision algorithms to develop new human mobile interaction (HMI) systems to overcome the limitations of the current HMI systems [2]. New 3D TVs and 3D displays can also benefit from human motion estimation. Basically, they display offset image frames that are filtered separately to the left and right eye. Estimating viewer’s head motion, the frames can be separately displayed to the viewer’s left and right eye to avoid wearing 3D glasses.

Intelligent environments describe physical environments in which information and communication technologies and sensor systems disappear as they become embedded into physical objects, infrastructures, and the surroundings in which we live, travel, and work [23]. Here, the goal is to allow computers to take part in activities never previously involved and allow people to interact with computers via gesture, voice, movement, and context.

Therefore, natural immersive interaction between users and computers is main characteristic of such applications. Here the question naturally arises: is there any interaction system that can meet users’ demands, and in addition, is easy to learn and use. For a long time, graphical user interfaces (GUIs) have been the dominant platform for human computer interaction. However, as computing becomes more widespread and ubiquitous, GUIs will not easily support the range of interactions necessary to satisfy users’ needs [24]. In order to accommodate a wider range of scenarios, tasks, users, and preferences, we need to move toward new interfaces that are more immersive, natural, intuitive, adaptive, and unobtrusive. With this motivation, the aim of a new generation of interfaces, Perceptual User Interfaces (PUIs), is to make human-computer interaction more like how people interact with each other and with the world [24]. Obviously, a detailed analysis of human motions and gestures is critically important to achieve perceptual user interfaces (see Fig. 1.3).
1.4 Research goal

The aims of this thesis are to develop computer vision theories for:

- Overcoming the limitations inherent in the current motion tracking systems.
- Providing higher motion resolution than the state of the art.
- Enabling the development of wearable human motion tracking and analysis systems.
- Directly recovering 3D human motion parameters from a sequence of range images.
- Creating natural and immersive forms of interaction between humans and technology.

1.5 Potential impact

- Biomedical applications. Many people around the world suffer from movement disorders. Only in the United States more than 40 million American are affected [25]. Based on a research conducted in Germany, there will be a dramatic increase in the number of people affected by most movement disorders between 2010 and 2050 [26]. For instance, the number of people who are diagnosed with Parkinson’s disease will increase up to 92% by
Designing accurate human motion analysis systems can be beneficial to diagnose and treat such a disease.

- **Entertainment.** In 2011 game console manufactures sold approximately 200 million game consoles over the world, and this number will increase in the future. As graphical features become more realistic, the users expect more natural means of interaction with game consoles. Hence human gesture recognition will have a long way ahead in this area.

- **Natural interaction.** Total smartphone sales in 2011 reached 472 million units and are estimated to rise to 980 million units in 2015 [27]. It is also expected that the number of tablets sold in 2011 will increase from 54.7 million to 79 million units in 2012 [28]. Because PC adoption in emerging markets is growing fast, it is estimated that there will be more than two billion PCs in use by 2015 [29]. These growing numbers reveal the huge impact of HMI and HCI in the near future. Developing new techniques to estimate and analyze the human motion will make a breakthrough in this field.

### 1.6 Thesis outline

Proposed techniques for capturing human motion are discussed in depth in Chapters 2 and 3. Then in Chapters 4 and 5, design criteria and system evaluations are demonstrated. Numerous potential applications for human motion analysis are reviewed in Chapter 6. A summary of contributions from the thesis is provided in Chapter 7. Finally, concluding remarks and future directions are summarized in Chapter 8.
Part II

DEVELOPING MOTION CAPTURE TECHNIQUES
Chapter 2

Head Pose Estimation

2.1 Introduction

Head motion is used to convey rich information in our daily lives. For instance, a person will point his head to indicate who the intended target in a conversation is. In a similar way in a discussion, head direction provides a nonverbal cue to a listener when to switch roles and begin speaking. In addition to the information resulted from deliberate head gestures, the visual focus of attention can be inferred by monitoring a person’s head. Visual attention is naturally linked with eye gaze estimation, i.e. the ability to estimate the direction and focus of a person’s eye.

Basically, head pose provides a rough indication of gaze that can be estimated in situations when the eyes are not visible (like low-resolution imagery, or in the presence of eye-occluding objects like sunglasses). When the eyes are visible, head pose becomes a requirement to accurately predict gaze direction. Knowledge about gaze direction can be deployed in various applications such as video conferencing and human-computer interfaces. In the context of video compression, robust 3D head pose estimation is substantially helpful tool to remove redundant information from video frames, and improve the level of data compression. Furthermore, face recognition systems can also benefit from robust and efficient head motion estimation technique.

In recent years there has been much research effort spent on creating new generation of user interfaces that enable natural, intuitive, and immersive interactions between humans and computers. As a consequence, moving from classical graphical user interfaces to natural perceptual user interfaces is undeniable. Therefore, understanding and analyzing the head motion is crucially important for delivering natural, unobtrusive and intuitive interactions. For instance, head pose estimation can enhance human computer interaction to a large extent. Controlling computer mouse and responding to pop-up dialog boxes using head movements are only
some existing examples that demonstrate how head pose estimation can change HCI systems.

Recall from the previous Chapter, 3D TVs and displays can also deliver intuitive and interactive experiences employing head pose estimation techniques. Since the major challenge in these systems is to transmit offset image frames to human eyes, head motion analysis could endow them with the capability to localize and track the eyes for displaying separate image frames to each eye.

2.2 Related work

3D head pose estimation is a challenging problem in the computer vision field owing to pose variations, illumination and complexity of the backgrounds. Conventional head pose estimation methods incorporate images taken by cameras as the input. Appearance template methods use image-based comparison metrics to match a new image of a head to a set of exemplars with corresponding pose label in order to find the most similar view [30] [31]. Some methods use the location of features such as the eyes, mouth, and nose tip to determine the head pose from their relative configuration [32] [33].

Tracking methods estimate the global pose change of the head from the observed movement between video frames [34] [35]. These methods involve extracting keypoints in the image, such as scale-invariant feature transform (SIFT) [22] to recover the motion from one frame to another frame. Though, these methods suffer from illumination changes. Range images, on the contrary, are well known to be robust against illumination variations in the environment and can be considered as a solution. In addition dealing with multiple users tracking where heads of different people overlap with each other or are occluded by other objects is still an issue [36] [37] [38].

RGB image based approaches encounter difficulties in retrieving the head. Employing depth information will substantially enhance the head retrieval since individual heads are discriminated from each other due to the knowledge of their corresponding depths. In the past few years, researchers have focused on using time-of-flight range cameras (TOF). They have proposed different algorithms to address the problem of pose estimation and human motion capture from range images [39] [40] [41] [42]. Although these methods have acceptable performance, they are limited in the sense that the all six DOF of the head motion cannot be recovered.
2.3 ACTIVE & PASSIVE MOTION CAPTURE

In this section the concepts of active and passive motion capture systems are clarified, and a technical comparison between these two methods is presented. Conventionally, vision-based human motion tracking systems place the camera in a particular point, where the camera can see the user. Thus, the user has to perform desired movements and gestures in the camera’s field of view. We address such a configuration as passive motion capture system. However, there is another way; we suggest mounting the camera on the human body and performing motion tracking. Therefore, the subject is not limited to be in the camera’s field of view. We refer to this system as active motion capture system. When using the passive configuration, certain issues must be considered. As mentioned in section 1.2, in some cases there is a need to use special markers to detect and track human body motion. Consequently, the system can fail due to the incorrect marker detection.

Other problems such as cluttered scene, human occlusion, scale variation and illumination can degrade the system performance. Nevertheless, the most essential drawback associated with the passive systems 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 [43]. But, we believe these challenges can easily be resolved employing active motion tracking. Since the camera is mounted on the user body, there is no need to detect special markers to track user motion. Instead, stable key points are extracted in the video frames. These

Figure 2.1: Top view of a head and a fixed camera. The head turns with angle $\theta$ causing a change in the resulted image. The amount of change depends on the camera location (A or B) [3].
CHAPTER 2. HEAD POSE ESTIMATION

points will be tracked in consecutive frames for estimating the human motion. In
the proposed approach SIFT [22] algorithm is used to detect key points. SIFT
features are scale invariant, and highly robust against illumination changes.

In addition, active motion tracking 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 [43]. In order to simplify the idea, consider a simple rotation around y-axis
as it is illustrated in Fig. 2.1. This figure shows a top view of an abstract human
head and a camera. Two possible configurations for human motion tracking are
presented, placing the camera at point A, in front of the user (the passive setup)
and mounting the camera on the head at point B (the active setup). As the user
turns with angle $\theta$, the horizontal change ($\Delta x$) in captured images is calculated for
both setups based on the perspective camera model. Let’s assume $\theta = 45^o$, then
for the passive motion tracking:

$$\Delta x_1 = f \frac{r_1}{\sqrt{2}r_2 - r_1}$$ (2.1)

and for the active motion tracking:

$$\Delta x_2 = f \frac{r_2}{r_2}$$ (2.2)

$$f \frac{r_1}{\sqrt{2}r_2 - r_1} \ll f \Rightarrow \Delta x_1 \ll \Delta x_2$$ (2.3)

For example, if $f = 100$, $r_1 = 15cm$, $r_2 = 80cm$, then the change for both cases
will be:

$$\Delta x_1 = \left( \frac{0.15}{\sqrt{2} \ast 0.8 - 0.15} \right) \ast 100 \approx 15.3 \text{ pixels}$$ (2.4)

$$\Delta x_2 = 100 \text{ pixels}$$ (2.5)

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.

2.4 Active head motion estimation

Fig. 2.2 depicts active motion tracking system overview. In this particular scenario,
we want to measure user’s head motion. A wearable camera is mounted on the
2.4. ACTIVE HEAD MOTION ESTIMATION

Figure 2.2: Active motion tracking system.

user’s ear. It should be realized that the camera can be either used to record the user’s head movements during daily life activities for offline analysis, or to provide live video frames for online analysis. As he turns his head, the video frames from the camera are fed to the system. Then stable interest points in the scene are extracted. These points are tracked in the next frame to find point correspondences. Afterwards, head motion parameters are recovered.

2.4.1 Head pose estimation

In order to analyze and estimate the head motion, stable key points have to be detected within the entire image. Among different feature detectors, SIFT is selected owing to its invariance to image transformation. Next, feature point correspondences are found between consecutive frames using pyramidal Lucas-Kanade optical flow algorithm [44]. This method is appropriate for fast motion tracking and has a low computational cost which is of our interest in real time applications. Two consecutive frames and corresponding key points are illustrated in Fig. 2.3. After finding point correspondences, a fundamental matrix for each image pair is computed using robust iterative RANSAC algorithm [45]. RANSAC algorithm is used to detect and remove the wrong matches and improve the performance. Running RANSAC algorithm, a candidate fundamental matrix is computed based on 8-point algorithm [46]. The fundamental matrix $F$ is the $3 \times 3$ matrix that satisfies the epipolar constraint

$$x_i^T F x_i = 0 \quad (2.6)$$

where $x_i$ and $x_i'$ are a set of image point correspondences. Each point correspondence provides one linear equation in the entries of $F$. Due to the fact that $F$ is defined up to a scale factor, $F$ can be computed from 8 point correspondences [46]. If the intrinsic parameters of the cameras are known, as they are in our case, the cameras are said to be calibrated. Then, a new matrix $E$ can be introduced by equation

$$E = K'^TFK \quad (2.7)$$
where the matrix $E$ is called the essential matrix, $K'$ and $K$ are $3 \times 3$ upper triangular calibration matrices holding intrinsic parameters of the cameras for two views. Once the essential matrix is known, the relative translation and rotation matrices, $t$ and $R$ can be recovered from it. Let the singular value decomposition of the essential matrix be

$$E \sim U \text{diag}(1, 1, 0) V^T$$  \hspace{1cm} (2.8)

where $U$ and $V$ are chosen such that $\det(U) > 0$ and $\det(V) > 0$ ($\sim$ denotes equality up to scale). If we define the matrix $D$ as:

$$D \equiv \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Then $t \sim t_u \equiv [u_{13} \ u_{23} \ u_{33}]^T$ and $R$ is equal to $R_a \equiv UDV^T$ or $R_b \equiv UD^T V^T$. If we assume that the first camera matrix is [$I$ | $0$] and $t \in [0, 1]$, there are then 4 possible configurations for second camera matrix: $P_1 \equiv [R_a \mid t_u]$, $P_2 \equiv [R_a \mid -t_u]$, $P_3 \equiv [R_b \mid t_u]$ and $P_4 \equiv [R_b \mid -t_u]$. One of these solutions corresponds to true configuration. In order to determine the true solution, one point is reconstructed using one of four possible configurations. If the reconstructed point is in front of both cameras, the solution corresponds to the true configuration [46]. Once the right configuration is obtained, the relative head motion between two consecutive frames is computed. For instance, in Fig. 2.3 the relative head ro-
tion between two consecutive images are $X = 1.6394, Y = -3.7986$, and $Z = -0.5870$ degree respectively.

### 2.5 Passive head motion estimation

This section addresses the problem of nonlinearity inherent in conventional 3D motion estimation techniques using only 2D RGB images. Then an innovative approach is presented based on the range images taken by Kinect to tackle this issue, which results in a passive head pose estimation method.

#### 2.5.1 3D linear & 2D nonlinear methods

Here the same notations used by Horn [47] 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. If we consider a moving camera in a static environment, then 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

$$V = -t - \omega \times r.$$  \hfill(2.9)

If we define the components of $t$ and $\omega$ as

$$t = (U, V, W)^T \quad \text{and} \quad \omega = (A, B, C)^T$$

we can rewrite the equation in component form as

$$\dot{X} = -U - BZ + CY$$ \hfill(2.10)

$$\dot{Y} = -V - CX + AZ$$ \hfill(2.11)

$$\dot{Z} = -W - AY + BX$$ \hfill(2.12)

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 [44]. Let $(x, y)$ denote the coordinate of
CHAPTER 2. HEAD POSE ESTIMATION

Figure 2.4: Microsoft Kinect: (A) laser projector, (B) RGB camera, (C), monochrome CMOS camera [4].

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 [47]:

\[
u = \frac{\dot{X}}{Z} - \frac{X\dot{Z}}{Z^2} = \left( -\frac{U}{Z} - B + Cy \right) - x\left( -\frac{W}{Z} - Ay + Bx \right), \quad (2.13)
\]

\[
v = \frac{\dot{Y}}{Z} - \frac{Y\dot{Z}}{Z^2} = \left( -\frac{V}{Z} - Cx + A \right) - y\left( -\frac{W}{Z} - Ay + Bx \right). \quad (2.14)
\]

The resultant equations for the optical flow are inversely proportional to the distance of \( P \) to the camera \((Z)\). Unlike the motion parameters (i.e. \( A, B, C, U, V, W \)) 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:

\[
x(UC - Wv + AW) + y(VC + Wu + BW) + xy(BU + VA) - y^2(CW + AU)
\]

\[-x^2(VB + CW) - V(B + u) + U(v - A) = 0. \quad (2.15)
\]

Here, the problem arises, since the final equation is nonlinear.

However, this issue can be simply resolved by acquiring the depth information from range images. In the following, a passive 3D linear method is proposed to directly estimate the head pose based on the range images obtained from Kinect.
2.5. PASSIVE HEAD MOTION ESTIMATION

Figure 2.5: Passive head pose estimation system [4].

2.5.2 Kinect

Kinect is a peripheral device for Microsoft Xbox 360. It can be applied to obtain depth estimations using a structured light pattern. The device consists of a multi-array microphone, a RGB camera, a monochrome CMOS camera, and an infrared laser projector (Fig. 2.4).

The laser projector produces a structured light pattern in the scene, which is imaged by the CMOS camera. The displacement of the CMOS camera relative to the laser projector results in computing the distance to objects in the scene using triangulation. The device is capable of outputting RGB, and range images with $640 \times 480$ pixels at 30 frames per second. Microsoft has released a non-commercial Kinect software development kit (SDK) [21] for Windows. It provides Kinect capabilities to developers who build applications with C++, C#, or Visual Basic by using Microsoft Visual Studio 2010. In addition, open source drivers in the form of the libfreenect [48] library are available and can be used to interface with the device. Approximate formula for converting Kinect depth map to metric distances are also available [49].

2.5.3 System description

This part presents an overview of the main steps in the proposed approach, which is demonstrated in Fig. 2.5. Given an input depth array, we first reduce noise and smooth the array for further process. Then a 3-stage head detection process is used to locate the user’s head. First, background subtraction is performed to isolate the foreground. Then, in order to find distinct objects, the foreground is passed through our algorithm. Finally, irrelevant candidate segments are discarded, and the user’s head is located. When the head is located in one frame, the system keeps track of it in next frames. Consequently, the head does not need to be detected again in coming frames. Once the head is segmented in two consecutive frames, the six DOF of the head motion can be recovered. Eventually, the head motion parameters can be used to facilitate different applications, such as human-computer interaction.
2.5.4 Multiple head detection and tracking

Head detection algorithm is composed of different steps, as shown in Fig. 2.6. After smoothing the depth array and reducing the noise, the raw depth values should be converted to metric values. The raw depth data will be converted to metric value between 0.6 and 6 meters according to the formula given by Stéphane Magnenat [49]. In the next step the depth array is subtracted from the background. It is assumed that the prior knowledge about the background is available. This can be considered as an initialization step, where the background depth array is extracted. A difference matrix is computed by comparing the original depth array with the background. If the difference is below a threshold, the pixel is set to zero otherwise it will be retained, resulting in a matrix containing the depth information of the foreground.

Then segmentation is performed through a depth-first algorithm. A pixel is in the same segment as its neighbor if the depth difference between them is less than a threshold (0.1-0.2 m). Any segment that contains fewer pixels than a particular number is considered as non-human and discarded. Given a segment the system also needs to locate the head. This is accomplished by finding the topmost pixel of the segment, estimating the height of the head, and finding the leftmost and rightmost pixels within a certain area belonging to the segment. These four pixels constitute the boundaries of the rectangle containing the head (Fig. 2.7).

In order to perform head tracking between frames, the mean x, y and depth values for each segment in one frame are stored and compared with those in the next frame. If a similar segment is found between frames, they are regarded as the same segment.

2.5.5 3D head pose estimation

The time-varying depth map from Kinect can be viewed as a function of the form \( Z(X,Y,t) \). Taking a full time derivative of \( Z \) via the chain rule, the following equation is obtained

\[
\frac{dZ}{dt} = \frac{\partial Z}{\partial X} \frac{dX}{dt} + \frac{\partial Z}{\partial Y} \frac{dY}{dt} + \frac{\partial Z}{\partial t}
\]  

(2.16)

This can be written in the form

\[
\dot{Z} = p\dot{X} + q\dot{Y} + Z_t
\]

The above equation will be called the depth rate constraint equation, where the three partial derivatives of \( Z \) are denoted by

\[ p = \frac{\partial Z}{\partial X}, \quad q = \frac{\partial Z}{\partial Y}, \quad \text{and} \quad Z_t = \frac{\partial Z}{\partial t} \]
2.5. PASSIVE HEAD MOTION ESTIMATION

![Diagram of head detection scheme](image)

**Figure 2.6:** *Head detection scheme* [4].
and the components of velocity of a point in the depth image are given by

\[
\dot{X} = \frac{dX}{dt}, \quad \dot{Y} = \frac{dY}{dt}, \quad \text{and} \quad \dot{Z} = \frac{dZ}{dt}.
\]

The values of the partial derivatives \( p \), \( q \), and \( Z_t \) can be estimated at each pixel in the depth map, while \( \dot{X} \), \( \dot{Y} \), and \( \dot{Z} \) are unknown.

There is one such an equation for every point in the segmented depth map corresponding to the head, so that if it contains \( n \) points, there are \( n \) equations in a total of \( 3n \) unknowns. The system of equations is extremely underconstrained and additional assumptions are necessary to provide a unique solution. In the above discussion no constraint on the motion of neighboring points was assumed, each point being able to move completely independently. Although, in most real motions, neighboring points within the head do have similar velocities. Horn and Harris [50] have shown that there is a way to increase the amount of constraint.

In analogy with a so-called direct method for recovering motion from an ordinary image sequence [51], we could assume that the sensor is rigid and that we have to recover the motion of the head relative to the sensor. In this case, there are only six degrees of freedom of motion to recover, so that the corresponding system of equations is now vastly overconstrained.

Let \( \mathbf{R} = (X, Y, Z)^T \) be a vector to a point on the head. If the head moves with instantaneous translational velocity \( \mathbf{t} \) and instantaneous rotational velocity \( \omega \) with respect to the sensor, then the point \( \mathbf{R} \) appears to move with a velocity

\[
\frac{d\mathbf{R}}{dt} = -\mathbf{t} - \omega \times \mathbf{R} \tag{2.17}
\]

with respect to the sensor [52]. The components of the velocity vectors are given by

\[
\mathbf{t} = \begin{pmatrix} U \\ V \\ W \end{pmatrix} \quad \text{and} \quad \omega = \begin{pmatrix} A \\ B \\ C \end{pmatrix}
\]
Rewriting the equation for the rate of change of $\mathbf{R}$ in component form yields

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

where the dots denote differentiation with respect to time. Substituting these expanded equations into the depth rate constraint equation itself yields

$$pU + qV - W + rA + sB + tC = Z_t$$

where

$$r = -Y - qZ, \quad s = X + pZ, \quad and \quad t = qX - pY$$

If there are $n$ pixels in the head area, the resulting $n$ equations can be written in a matrix form as

$$A\mathbf{x} = \mathbf{b}$$

where

$$A = \begin{bmatrix} p_1 & q_1 & -1 & r_1 & s_1 & t_1 \\ p_2 & q_2 & -1 & r_2 & s_2 & t_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_n & q_n & -1 & r_n & s_n & t_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (Z_1)_1 \\ (Z_1)_2 \\ \vdots \\ (Z_1)_n \end{bmatrix}$$

or

$$A\mathbf{x} = \mathbf{b}.$$
CHAPTER 2. HEAD POSE ESTIMATION
Chapter 3

Hand Motion Estimation

3.1 Introduction

Presently, owing to the large influx of computers in our daily lives, human computer interaction has become substantially important. Traditionally, keyboard and mouse have played the main role in HCI. However, with the rapid progress of computing, communication, and display technologies, such interfaces may become a bottleneck in applications that rely on heavy interaction of user with a machine due to the unnaturalness of the interaction. This limitation has become even more apparent with the emergence of novel display technology such as virtual reality. Thence, researchers have recently attempted to eliminate this HCI
bottleneck by developing more natural ways of interaction. With this motivation automatic speech recognition and human gesture recognition have been topics of research for decades. Human gestures may be defined as physical movements of hands, arms, face, and body with the intent to convey information or meaning. In this thesis human computer interactions through hand gestures will be considered. Human hand gestures are a means of non-verbal interaction among people. They range from interactions with objects (manipulative gestures) to more complex ones that express feelings or provide a form of communication amongst humans. Several taxonomies have been suggested in the literature that deals with psychological aspects of gestures. The one that fits well with the context of HCI was developed by Quek [53] [54](see Fig. 3.1).

3.2 Previous implementation

To exploit the use of gestures in HCI it is necessary to put an accurate visual interpretation on human gestures. Earliest attempts to solve this problem have employed mechanical glove-based devices that directly measure the hand pose and/or hand joint angles [55] [56] [57]. In glove-based approaches, user is required to wear a cumbersome device, and generally carry a load of cables that connect the device to a computer. This hampers the ease and naturalness of interaction between user and computer. Additionally, the glove-based gestural interfaces are not cost effective, the matter that put them out of reach for general use.

The aforementioned drawbacks in glove-based systems can be overcome using vision-based interaction techniques. This approach suggests using a set of video cameras and computer vision techniques to interpret gestures and is the focus of this project. Computer vision based techniques are relatively cost-effective methods to acquire and interpret human hand gestures, and minimally obtrusive to participants.

Most of the vision-based systems rely on extracting feature sets for the purpose of hand gesture recognition. Hand features can be derived using the following three approaches:

- Model based approaches

Model based approaches attempt to infer the pose of the palm and the joint angles [58] [59] [60] [61]. Such an approach would be ideal for realistic interactions in virtual environments. Generally, the approach consists of searching for the kinematic parameters that brings the 2D projection of a 3D model of hand into correspondence with an edge-based image of a hand. This approach cannot handle the inevitable self-occlusion of the hand. More recent efforts have reformulated
3.2. PREVIOUS IMPLEMENTATION

the problem within a Bayesian (probabilistic) framework [62] [63]. Bayesian approaches allow for the pooling of multiple sources of information (e.g. system dynamics, prior observations) to arrive at both an optimal estimate of the parameters and a probability distribution of the parameter space to guide future search for parameters.

A common problem with the model-based approaches is the problem of the feature extraction (i.e. edges). The human hand itself is rather texture less and does not provide many reliable edges internally. The edges are usually extracted from the occluding boundaries. In order to facilitate extraction and unambiguous correspondence of edges with models, these methods require homogeneous backgrounds and high contrast backgrounds relative to the hand.

- View based approaches

View-based approaches, also referred to as appearance-based approaches, model the hand by a collection of 2D intensity images, i.e. gestures are modeled as a sequence of views [64] [65] [66] [67]. Currently, eigenspace approaches represent the state-of-the-art for view based approaches. The eigenspace approach provides an efficient representation of a large set of high-dimensional points using a small set of basis vectors. Given the success in face recognition, many have applied the eigenspace approach to hand gestures [64] [67]. For a small set of gestures this approach may be sufficient. With a large gesture vocabulary the space of views is large, this poses a problem for collecting adequate training sets and more seriously the compactness in the subspace required for efficient processing may be lost.

- Low-level features

Many approaches have utilized the extraction of low-level image measurements that are fairly robust to noise and can be extracted quickly. Low-level features that have been proposed in the literature include: the centroid of the hand region [68],
3.3 Vision-based gesture detection & tracking

In this part the proposed 3D camera-based interaction approach is presented. As user moves his/her hand gesture in the camera’s field of view behind the mobile device, the device captures a sequence of images. Then this input will be processed in gesture detection block. As a result, the user gesture will be detected and localized. Afterwards, stable features in each image are extracted to compute the relative rotation and translation between two frames. Finally, this information can be used to facilitate human mobile interaction and manipulation of virtual objects on the screen. Fig. 3.2 depicts the system overview for proposed approach.

3.3.1 Gesture detection

The first step to interact with the mobile phone is to detect and localize the user gesture. The gesture detection algorithm relies on the Rotational Symmetry [71] patterns. Rotational Symmetries are specific curvature patterns detected from local orientation. The main idea behind the rotational symmetries theory is to use local orientation to detect complex curvatures in double-angle representation [71]. Using a set of complex filters on the orientation image will result in detection of number of features in different orders, such as curvatures, circular and star patterns [71]. Fig. 3.3 illustrates three different orders of rotational symmetries.

In the suggested method, the gesture detection system takes the advantage of the rotational symmetries to localize the user gesture in image sequences which lead us to differentiate between fingers and other features even in complicated backgrounds. Since the natural and frequently used gesture to manipulate objects in 3D space is similar to Fig. 3.4(a), this model can satisfy our expectations for
3.4. GESTURE MOTION ESTIMATION

By localizing the hand gesture, a region of interest is defined around it. Then, the same algorithm described in section 2.4.1 is employed to recover hand gesture different applications. We aim to design our gesture detection system to detect and localize this particular gesture in image sequences for further processing. Our experiments based on different test images of different scales and backgrounds reveal that the user gesture substantially responds to the second order rotational symmetry patterns (circular patterns). Thus, our gesture detection system is designed to detect circular pattern in the input image. The double-angle representation of a given image can be computed as:

$$z(x) = (f_x(x) + if_y(x))^2 = f_x^2(x) - f_y^2(x) + 2i f_x(x)f_y(x)$$  \hspace{1cm} (3.1)

where local orientation is defined as, $$f(x) = (f_x(x) f_y(x))^T$$. Eventually, to detect the 2nd order rotational symmetries in an image, the double-angle image should be correlated with the complex filter $$a(x)b_2(x)$$, where $$b_2(x) = e^{i\phi}$$ is the 2nd order symmetry basis function, and $$a(x)$$ is a weight window for the basis function. In each local region in an image we compute the scalar product

$$S_2 = \langle ab_2, z \rangle$$  \hspace{1cm} (3.2)

High magnitudes in the result $$S_2$$ indicate the higher probability of 2nd order rotational symmetries patterns in the image. Our observation shows that searching for the second order rotational symmetries in image frames with suitable filter size will result in high probability of responses of user gesture in different scales. Consequently, this will result in a proper localization of the user gesture (Fig. 3.4(b)).

3.4 Gesture motion estimation

By localizing the hand gesture, a region of interest is defined around it. Then, the same algorithm described in section 2.4.1 is employed to recover hand gesture
CHAPTER 3. HAND MOTION ESTIMATION

Figure 3.5: Feature matching in two consecutive frames where 54 point correspondences are detected [2].

motion. Stable features in the scene are detected and tracked in the next frame to find match correspondences. Given point correspondences, fundamental and essential matrices are computed, and finally hand gesture motion is estimated. For instance, in Fig. 3.5 the relative rotation between two consecutive images are $X = 1.6892$, $Y = -0.4269$, and $Z = -1.5406$ degree.

During hand gesture motion estimation process, it is presumed that the camera motion is limited to the smooth user’s hand shaking; otherwise the camera motion should be distinguished from the gesture motion for more accurate motion estimation.
Part III

EVALUATING MOTION CAPTURE SYSTEMS
Chapter 4

Evaluation of Head Motion Estimation Systems

4.1 Introduction

For a head pose estimation system to be of general use, it should be invariant to identity, have sufficient range of allowed motion, require no manual intervention, and should be easily deployed on conventional hardware. To satisfy the majority of applications, the following design criteria are proposed in [72] for head pose estimation systems.

4.2 Design criteria

- **Accurate**: The system should provide a reasonable estimate of pose with a mean absolute error of $5^\circ$ or less.

- **Monocular**: The system should be able to estimate head pose from a single camera. Although accuracy might be improved by stereo or multi-view imagery, this should not be a requirement for the system to operate.

- **Autonomous**: There should be no expectation of manual initialization, detection, or localization.

- **Multi-Person**: The system should be able to estimate the pose of multiple people in one image.

- **Identity & Lighting Invariant**: The system must work across all identities with the dynamic lighting found in many environments.
CHAPTER 4. EVALUATION OF HEAD MOTION ESTIMATION SYSTEMS

- **Resolution Independent**: The system should apply to near-field and far-field images with both high and low resolution.

- **Full Range of Head Motion**: The methods should be able to provide a smooth, continuous estimate of head rotations, even when the face is pointed away from the camera.

- **Real-Time**: The system should be able to estimate a continuous range of head orientation with fast (30fps or faster) operation.

4.3 System evaluation

Although no single system has met all of the design criteria, we aim to fulfill these criteria. Based on the design criteria, two proposed systems are evaluated, which are explained in the following.

4.3.1 Passive head motion estimation

The passive 3D head pose estimation method has been implemented and examined with a quad core Intel i7 at 3.4 GHz. This approach is fast enough to operate in real time applications. The computational time for head detection block is about 15-25 ms and 10-15 ms for 3D pose estimation method. Consequently, the total processing time is about 25-40 ms which approximately corresponds to more than 25 frames per second. The head detection algorithm is tested on a set of 200 range images and the correct detection rate is almost 96%. We also tried to implement OpenCV’s Haar feature based face detection [73], using the Kinect’s RGB camera, to facilitate the head detection process, but two major problems was faced. First, this method is computationally expensive and takes about 45-70 ms to detect the human face. Second, it turned out that the user is limited to perform small head rotations to keep his/her face in front of the camera to have acceptable face detection rate. In the other words, if the user rotates his/her head more than a particular angle, which is natural in most real applications, Haar feature based face detector fails to locate the face.

Since there is no Ground Truth available, an experiment is designed to evaluate the system performance. In this experiment the user’s head is detected and located in the range images by the system and then the six DOF of the head motion are recovered and used to manipulate a 3D model on the computer screen. As it is shown in Fig. 4.1, the position and orientation of the cubes is updated whenever the user moves his head. Our experiments revealed that the effective distance from the sensor ranges from 0.6 up to 6 meters.
Figure 4.1: Experimental results. The 3D head motion parameters are estimated to update the position and orientation of the 3D model. First row is the initial position. Next three rows show the rotation around X, Y, and Z axes respectively. The last three rows illustrate the translation in X, Y, and Z [4].
4.3.2 Active head motion estimation

Several tests have been performed to report the angular accuracy of the active tracking system. We developed an electronic measuring device to evaluate our proposed system (see Fig. 4.2). The electronic device outputs are used as Ground Truth to evaluate the active motion tracking system. 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 motor can be operated by C codes through the control board. It can move in two different directions with specified speed, and its true rotation value (Ground Truth) is indicated on the protractor. As the servo turns, the captured image frames will be processed and the camera rotation will be
estimated by the active tracking system. Then the system outputs are compared to Ground Truth to evaluate the system. Three different setups are used to test the system around X, Y, and Z-axis (Fig. 4.2 a, b, and c). The tests have been carried out on an HP machine with an Intel core 2 Duo, 2.93 GHz processor. A Logitech Webcam 905 was used with a resolution of 640X480. Depending on the image content, 280 to 500 SIFT interest points were extracted per image. The system continuously measured the camera motion at the rate of 25 Hz by analyzing interest points. The camera is rotated from 0 to 40 degree around three axes separately, and the mean absolute error is calculated for each turn. The system evaluation was repeated for 50 times and the results are presented in Table 4.1. The error increases as the camera rotates, as it was expected. When the camera turns around X-axis, the number of missed interest points is larger than when rotating around Y and Z-axis. As a result, the error is slightly larger in X-axis. However, our system is more accurate and robust compared to most of the current vision-based tracking systems, which aim to provide reasonable motion estimation with a mean absolute error of $5^\circ$ or less [72]. Taking advantage of the active tracking system, we obtained mean absolute errors of $0.5779^\circ$, $0.3047^\circ$, and $0.2449^\circ$ for small rotations ($5^\circ$), and $2.4038^\circ$, $1.4485^\circ$, and $0.7227^\circ$ for large motions ($40^\circ$) around X, Y and Z-axis respectively. Moreover, another experiment was developed to show the system usability (Fig. 4.3). Mounting the camera on the user’s head, the system estimates the user head motion and records the data. Motion parameters are applied to control a 3D model on the computer screen to visualize the user head motion.

![Active motion tracking demo](image)

**Figure 4.3:** *Active motion tracking demo. As the user turns his head, the motion parameters are estimated and used to change the 3D model on the computer screen* [3].
### Table 4.1: System evaluation data sheet. Data in the left column are actual rotation angles and the other columns are MAE (Mean Absolute Error) around X, Y, and Z axes (in degree).

<table>
<thead>
<tr>
<th>Rotation angle</th>
<th>MAE(X-axis)</th>
<th>MAE(Y-axis)</th>
<th>MAE(Z-axis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.5779</td>
<td>0.3047</td>
<td>0.2449</td>
</tr>
<tr>
<td>10</td>
<td>0.6647</td>
<td>0.3914</td>
<td>0.2755</td>
</tr>
<tr>
<td>15</td>
<td>0.7722</td>
<td>0.4857</td>
<td>0.3697</td>
</tr>
<tr>
<td>20</td>
<td>0.9316</td>
<td>0.5915</td>
<td>0.4515</td>
</tr>
<tr>
<td>25</td>
<td>1.5077</td>
<td>0.6289</td>
<td>0.4570</td>
</tr>
<tr>
<td>30</td>
<td>1.7645</td>
<td>0.6994</td>
<td>0.4782</td>
</tr>
<tr>
<td>35</td>
<td>1.9715</td>
<td>0.9814</td>
<td>0.6491</td>
</tr>
<tr>
<td>40</td>
<td>2.4038</td>
<td>1.4485</td>
<td>0.7227</td>
</tr>
</tbody>
</table>

### 4.4 Conclusion

Here, according to the design criteria introduced in section 4.2, a technical comparison between two proposed methods for estimating head pose is presented. Table 4.2 illustrates the comparison.

We cannot judge how accurate the passive system is because there is no Ground Truth available to evaluate it. Nevertheless, both systems are monocular. The active method is autonomous, although the passive one needs the initialization step. They can estimate the head pose of multiple users in the scene, and both

<table>
<thead>
<tr>
<th>Design Criteria</th>
<th>Head Pose Estimation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Active</td>
</tr>
<tr>
<td>Accuracy (in 40°)</td>
<td>MAE(X-axis) = 2.4038°</td>
</tr>
<tr>
<td></td>
<td>MAE(Y-axis) = 1.4485°</td>
</tr>
<tr>
<td></td>
<td>MAE(Z-axis) = 0.7227°</td>
</tr>
<tr>
<td>Monocular</td>
<td>√</td>
</tr>
<tr>
<td>Autonomy</td>
<td>√</td>
</tr>
<tr>
<td>Multiuser</td>
<td>√</td>
</tr>
<tr>
<td>Identity &amp; Lighting</td>
<td>Invariant</td>
</tr>
<tr>
<td>Resolution</td>
<td>No Limitation</td>
</tr>
<tr>
<td>Rang of Head Motion</td>
<td>Full Range</td>
</tr>
<tr>
<td>Processing Time</td>
<td>25 fps</td>
</tr>
</tbody>
</table>

Table 4.2: Technical comparison between proposed head pose estimation methods.
of them are identity and lighting invariant. The active setup has no resolution restriction, whereas the passive system works in the range of 0.6-6.0 meters from the sensor. While the passive setup has limited freedom to rotate around the Z-axis, the active system recovers full range of head motion. Eventually, both systems are fast enough to process 25 frames per second.
Chapter 5

Evaluation of Hand Gesture Recognition System

5.1 Introduction

Although vision-based gestural interaction is a challenging interdisciplinary research area, people desire off-the shelf, universal vision-based interfaces that can be put to use in any new application. To make a system that can recognize and interpret human gestures, there are some requirements which should be met by the system.

5.2 Design criteria

- **Effectiveness**: In the real-world, visual information could be very rich, noisy, and incomplete, due to illumination variation, clutter and dynamic backgrounds, occlusion, etc. The system should be robust against all these factors.

- **Efficiency**: Human computer interaction systems require real-time interfaces. So the system should be able to recognize human gestures with fast operation.

- **User satisfaction**: Vision-based gestural interaction systems should deliver natural, intuitive, and immersive experience to users.
CHAPTER 5. EVALUATION OF HAND GESTURE RECOGNITION SYSTEM

Figure 5.1: (a) System performance in gesture tracking. (b) Error of the tracking in a sequence of images [5].

5.3 System Evaluation

The main objective here is to demonstrate whether or not the vision-based gestural interaction system meets the above-mentioned criteria. Real-time experiments in complex environments are conducted to evaluate user freedom for interaction with, and manipulation of virtual objects. Our experiments revealed that in order to have a robust detection and consequently tracking, our system should be scale and rotation-invariant to the user gesture. As a matter of fact, for a particular gesture behind the mobile phone’s camera, users have freedom to move in a reasonable distance. Moreover, depending on the application, they are free to rotate in different angles.

Our observation indicates that the effective interaction happens in the area between 15-25cm from the camera. Interaction in the area beyond 25cm does not seem to be convenient for users. Clearly, for distances below 15cm, gesture occupies a large area on the screen and degrades the interaction. Fig. 5.1(a) illustrates the system performance in the tracking of the particular curve on a complex back-
5.3. SYSTEM EVALUATION

In this example the user is asked to follow the predefined curve drawn on the screen. Circles mark the position of the detected gesture corresponding to each image frame. The error in the tracking of the original curve for 230 frames is plotted in Fig. 5.1(b). The mean value of the error (6.59 pixels) shows the slight difference between the original curve and the one plotted by the tracked gesture which is quite satisfying. After localizing the user gesture in the frames, the relative rotation and translation between image pairs will be computed. Fig. 5.2 shows an experiment where the goal is to estimate the gesture motion and apply it to a teapot. The teapot’s position and orientation will be updated according to the gesture motion.

Figure 5.2: Relative rotation and translation between image sequences. (a) The teapot should rotate in the direction specified by the arrow while the user gesture rotates in the same direction. (b) Consecutive frames (top), feature matching between two images (middle), and applying user gesture motion to the teapot (bottom) [2].
CHAPTER 5. EVALUATION OF HAND GESTURE RECOGNITION SYSTEM
Part IV

APPLICATION DOMAINS
Chapter 6

Applications

The potential applications of human motion analysis can be categorized into three major domains: surveillance, control, and analysis. The surveillance area covers applications where one or more subjects are being tracked and monitored for special actions. The control area relates to applications where the captured motion is used to provide controlling functionalities, and the third application area is concerned with the detailed analysis of the captured motion data.

Considering this classification, a few application areas are presented, in which human motion estimation has and will continue to have a profound impact.

6.1 Application areas

- **Video Surveillance**: A classic example is the bank surveillance system [74], where human motion estimation provides system controllers with the ability to closely monitor the activities of employees and customers to evaluate whether they may be about to commit a crime.

- **Intelligent Environments**: Head pose estimation systems will play a key role in the creation of intelligent environments. Already there has been a huge interest in smart rooms that monitor the occupants and use head pose to measure their activities and visual focus of attention [75] [76] [77]. Head pose endows these systems with the ability to determine who is speaking to whom, and to provide the information needed to analyze the non-verbal gestures of the meeting participants.

- **Virtual Reality**: Human gestures for virtual and augmented reality applications have experienced one of the greatest levels of uptake in computing. Virtual reality interactions use gestures to enable realistic manipulations of
CHAPTER 6. APPLICATIONS

virtual objects using users hands, for 3D display interactions [78] or 2D displays that simulate 3D interactions [79].

- **Robotics and Telepresence**: Telepresence and telerobotic applications are typically situated within the domain of space exploration and military-based research projects. The gestures used to interact with and control robots are similar to fully-immersed virtual reality interactions, however the worlds are often real, presenting the operator with video feed from cameras located on the robot [80]. Here, gestures can control a robot’s hand and arm movements to reach for and manipulate actual objects in the world.

- **Desktop and Mobile-device Applications**: In desktop computing, gestures can provide an alternative interaction to the mouse and keyboard [81]. Many gestures for desktop computing tasks involve manipulating graphics, or annotating and editing documents using pen-based gestures [82]. Head pose estimation could enable breakthrough interfaces for computing. Some existing examples include systems that allow a user to control a computer mouse using his head pose movements [83], respond to pop-up dialog boxes with head nods and shakes [84], or use head gestures to interact with embodied agents [85].

- **Entertainment**: Tracking and analyzing human body movements and gestures can enrich experience of computer games as well. Freeman et al. [86] tracked a player’s hand or body position to control movement and orientation of interactive game objects such as cars. Konrad et al. [87] used gestures to control the movement of avatars in a virtual world. Recently, game console manufactures have provided new devices that allow players to interact with games using motion, and also sound. Sony has introduced the PLAYSTATIONEYE, a camera that tracks hand movements for interactive games [88], and Microsoft has released Kinect for Xbox 360 which can detect and track human gestures to control and manipulate interactive game objects [21].

- **Medical Applications**: Movement disorder clinics and health care systems can also benefit from human motion analysis. Movement disorders affect speed, quality, and ease of human movements, and in many cases, surgical intervention is the best possible treatment for such a disease. Vision-based motion capture systems endow clinicians to observe patient’s improvements after surgical operation [15] [16] [17].

- **Automotive Safety**: Head pose estimation will have a profound impact on the future of automotive safety. An automobile driver is fundamentally limited by the field-of-view that one can observe at any one time. When one fails to
notice a change to his environment, there is an increased potential for a life-threatening collision that could be mitigated if the driver was alerted to an unseen hazard. As evidence to this effect, a recent comprehensive survey on automotive collisions demonstrated a driver was 31% less likely to cause an injury-related collision when there was one or more passengers [89]. Consequently, there is great interest in driver assistance systems that act as virtual passengers, using the driver’s head pose as a cue to visual focus of attention and mental state [90] [91] [92].
Part V

CONTRIBUTIONS, DISCUSSIONS, AND FUTURE DIRECTIONS
Chapter 7

Outline and Summary of Contributions

7.1 Summary of contributed papers

Paper I
Farid Abedan Kondori, Li Liu, “3D Active Human Motion Estimation for Biomedical Applications,” accepted in World Congress on Medical Physics and Biomedical Engineering (WC2012), Beijing, China, 26-31 May 2012.

Movement disorders prevent many people from enjoying their daily lives. As with other diseases, diagnosis and analysis are key issues in treating such disorders. Vision-based motion capture systems are helpful tools for accomplishing this task. However, classical motion tracking systems suffer from several limitations. First, they are not cost-effective. Second, these systems cannot detect minute motions accurately. Finally, they are spatially limited to the lab environment where the system is installed. In this project, we propose an innovative solution to solve the aforementioned issues. Mounting the camera on human body, we build a convenient, low-cost motion capture system that can be used by the patient in daily-life activities. We refer to this system as active motion capture, which is not confined to the lab environment. Real-time experiments in our lab revealed the robustness and accuracy of the system.

Paper II
Head pose estimation plays an essential role for bridging the information gap between humans and computers. Conventional head pose estimation methods are mostly based on the RGB images. However, accurate and robust head pose estimation is often problematic. In this paper, we present an algorithm for recovering the six degrees of freedom (DOF) of motion of a head from a sequence of range images taken by Kinect. The proposed algorithm utilizes a least-squares minimization of the difference between the measured rate of change of depth at a point and the rate predicted by the depth rate constraint equation. We segment the human head from its surroundings and background, and then we estimate the head motion. Our system has the capability to recover the head motion of multiple people in one image. The proposed system is evaluated in our lab and presents superior results.

**Paper III**

Designing a robust hand gesture detection system using a single camera independent of lighting conditions or camera quality is still a challenging issue in the field of computer vision. A common method for gesture detection is marker-based approach. Most of the augmented reality applications are based on marked gloves for accurate and reliable fingertip tracking. However, in marker-based methods users have to wear special inconvenient markers. What we present in this paper is simply based on low-level operators, detecting natural features without any effort to have an intelligent system. We propose a way to take advantage of lines, curvatures, circular patterns and in general rotational symmetries associated with the model of the human fingers which leads to the detection of the human gesture.

**Paper IV**

New mobile devices are equipped with integrated cameras and large displays which make the interaction with device easier and more efficient. Although most of the previous works on interaction between humans and mobile devices are based on 2D touch-screen displays, camera-based interaction opens a new way to manipulate in 3D space behind the device in the camera’s field of view. This paper
suggests the use of particular patterns from local orientation of input image called rotational symmetries to detect and localize human gesture. Relative rotation and translation of human gesture between consecutive frames are estimated by means of extracting stable features. Consequently, this information can be used to facilitate the 3D manipulation of virtual objects in various applications in mobile devices.

7.2 Summary of contributions

The contributions from this thesis can be divided into two areas. The key contributions to each area are presented here.

3D head pose estimation. The contributions to this area are from the papers I and II.

- Developing active motion capture system by mounting cameras on the human body, to provide higher resolution, and to overcome the problems of the existing human motion analysis systems.

- 3D linear head pose estimation method for directly recovering head motion parameters from a sequence of range images.

Hand gesture motion estimation. The contributions to this area are from the papers III and IV.

- Hand gesture detection and tracking algorithm based on the very low-level operations with the minimum level of intelligence.

- Hand gesture motion recovery by means of extracting stable features in the scene to develop new human mobile interaction systems.
CHAPTER 7. OUTLINE AND SUMMARY OF CONTRIBUTIONS
Chapter 8

Concluding Remarks & Future Directions

In the previous chapters current human motion capture systems have been reviewed (Chapter 1), active and passive head pose estimation methods have been presented (Chapter 2), hand gesture recognition and tracking scheme has been discussed (Chapter 3), and experimental results have been demonstrated (Chapter 4 and 5). In this chapter, the concluding remarks based on the four research studies conducted by the author and colleagues are summarized.

8.1 Conclusions

- We have proposed an active approach to capture human head motion. Mounting the camera on the head, we have built a convenient, low cost motion capture system that can be used by users even in daily-life activities. Scale changes, human occlusion, and cluttered scenes do not affect the system performance any longer. Additionally, active motion capture system is not confined to the lab environments, and provides high resolution and accuracy. However, detecting sufficient number of interest points in the scene is the only requirement for the proposed method.

- We have developed a passive head pose estimation system based on the depth information from range images. Our method is able to linearly estimate the 3D head pose of multiple persons in one image. Moreover our algorithm is identity and lighting invariant and works across all identities with the dynamic lighting condition. It can apply to near-field and far-field images with both high and low resolutions as well.
Although the experiments show the robustness and efficiency of the system, our method has high dependency on accurate head detection, which implies that if user is wearing a strange shape hat, or unusual hair style, it probably will degrade the head detection performance. The integration of RGB images into depth arrays can improve the head detection algorithm.

- Despite the fact that the passive approach is more robust against the lighting variations than the active approach, a few limitations are associated with the passive method in comparison with the active motion analysis system. First, the passive system needs initialization step. The head has to be detected and localized to estimate the head motion. Thus, wrong head detection and localization would affect the system performance. Second, the passive method is spatially limited. The user must be placed in front of the sensor, with the result that analyzing user movements during daily life activities is very unlikely. Last but certainly not least, small head movements cannot be detected accurately by the passive approach. Higher resolution in the active system results in estimating small movements accurately.

- We have also presented a novel approach for hand gesture motion estimation. Our detection algorithm can estimate the position of hand gesture in consecutive frames. This algorithm is computationally efficient and relies on low level operations, detecting natural features without any effort to have an intelligent system. Then the relative hand motion between two consecutive frames is estimated, which can be used to facilitate a diverse set of applications.

8.2 Future directions

The research presented in this thesis still leaves some open questions concerning improving the vision-based motion capture system, and implementing it in other related areas.

- Future research

There are two areas where further research is planned. First, we plan to apply data mining to the active motion capture system, to develop new vision-based theories. Second, improving the gesture recognition system to detect and recognize a wide range of hand gestures.
8.2. FUTURE DIRECTIONS

- Future implementations

There are numerous 3D visualization systems which can benefit from human motion estimation. In our lab, we have planned to implement 3D head motion analysis to these systems to enhance the quality of 3D experience. We also have a vision to take the advantage of head motion estimation to make a major breakthrough in the 3D displays in the future.
Bibliography


BIBLIOGRAPHY


