VISION-BASED THREE DIMENSIONAL HAND INTERACTION IN MARKERLESS AUGMENTED REALITY ENVIRONMENT

NG KAH PIN

UNIVERSITI SAINS MALAYSIA

2015

VISION-BASED THREE DIMENSIONAL HAND INTERACTION IN MARKERLESS AUGMENTED REALITY ENVIRONMENT

by

NG KAH PIN

Thesis submitted in fulfillment of
the requirements for the Degree of
Master of Science

May 2015

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Ms. Tan Guat Yew for the continuous support throughout my study and research. Her invaluable guidance, motivation and patience are greatly appreciated.

I would also wish to thank the Dean, lecturers and staff of School of Mathematical Sciences and Institute of Postgraduate Studies, Universiti Sains Malaysia for providing a conducive environment for my research and their generous help in various ways for the completion of my thesis.

My heartfelt appreciation goes to my family, for their unconditionally love, care and support at all times. To my friends, thank you for the encouragement and always being there for me.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

K. P. Ng, G. Y. Tan, Y. L. Iman, "Overview of Augmented Reality Tools," in *Proc. of Seminar Kebangsaan Aplikasi Sains & Matematik,* 2010, pp. 123-128.

K. P. Ng, G. Y. Tan, "ARTransform: Visualization of Three Dimensional Geometric Transformations in Augmented Reality Environment," in *Proc. of The International Conference on Computer Graphics and Virtual Reality*, 2011, pp. 145-149.

K. P. Ng, G.Y. Tan, "Deployment of Augmented Reality for Three Dimensional Viewing Analysis", *2nd Symposium of USM Fellowship 2011 (USMFS11)*, Penang, 23-24 Nov 2011.

K. P. Ng, G. Y. Tan, Y. P. Wong, "Vision-Based Hand Detection for Registration of Virtual Objects in Augmented Reality", *International Journal of Future Computer and Communication*, vol. 2, no. 5, pp. 423-427, 2013.

INTERAKSI TANGAN TIGA-MATRA BERASASKAN PENGLIHATAN KOMPUTER DALAM PERSEKITARAN REALITI TAMBAHAN TANPA PENANDA

ABSTRAK

Kemunculan realiti tambahan membolehkan objek maya untuk wujud bersama dengan dunia sebenar dan ini memberi kaedah baru untuk berinteraksi dengan objek maya. Sistem realiti tambahan memerlukan penunjuk tertentu, seperti penanda untuk menentukan bagaimana objek maya wujud dalam dunia sebenar. Penunjuk tertentu mesti diperolehi untuk menggunakan sistem realiti tambahan, tetapi susah untuk seseorang mempunyai penunjuk tersebut pada bila-bila masa. Tangan manusia, yang merupakan sebahagian dari badan manusia dapat menyelesaikan masalah ini. Selain itu, tangan boleh digunakan untuk berinteraksi dengan objek maya dalam dunia realiti tambahan. Tesis ini membentangkan sebuah sistem realiti tambahan yang menggunakan tangan terbuka untuk pendaftaran objek maya dalam persekitaran sebenar dan membolehkan pengguna untuk menggunakan tangan yang satu lagi untuk berinteraksi dengan objek maya yang ditambahkan dalam tiga-matra. Untuk menggunakan tangan untuk pendaftaran dan interaksi dalam realiti tambahan, postur dan isyarat tangan pengguna perlu dikesan. Algorithma penglihatan komputer digunakan untuk mengesan tangan tanpa bantuan penanda atau peranti-peranti yang lain seperti peranti mekanikal dan peranti magnet. Kamera stereo digunakan untuk mendapatkan imej video, supaya maklumat kedalaman tangan dapat diperolehi dengan menggunakan pendekatan penglihatan stereo. Segmentasi dengan warna kulit digunakan untuk mensegmen rantau tangan dari imej. Ciri-ciri tangan kemudiannya digunakan untuk mengesan postur tangan yang berlainan. Pusat tapak tangan dan

hujung jari tangan terbuka akan dikesan dalam masa nyata tetapi bukannya penanda. Dengan maklumat kedalaman yang diperolehi, kedudukan tiga-matra ciri-ciri tangan ini kemudiannya digunakan untuk menganggarkan posisi kamera berasaskan tangan. Posisi ini mempunyai enam darjah kebebasan dan akan digunakan untuk menambah objek maya di atas tapak tangan dengan tepat. Beberapa isyarat tangan digunakan untuk memanipulasi objek maya. Sistem realiti tambahan yang dikenalkan membenarkan pengguna melihat objek maya yang ditambahkan dari sudut pandangan yang berbeza dengan hanya menggerakkan tangannya. Di samping itu, pengguna juga dapat menggunakan tangan yang lain untuk memilih dan menggerakkan objek maya. Sistem ini dapat mencapai kadar sebanyak 12 fps. Postur tangan dapat dikesan dengan tepat. Ketepatan untuk penganggaran posisi dan orientasi kamera bukan sangat tinggi, tetapi dapat diterima untuk menambahkan objek maya dalam persekitaran sebenar.

VISION-BASED THREE DIMENSIONAL HAND INTERACTION IN MARKERLESS AUGMENTED REALITY ENVIRONMENT

ABSTRACT

The advent of augmented reality (AR) enables virtual objects to be superimposed on the real world and provides a new way to interact with the virtual objects. AR system requires an indicator to determine for how the virtual objects aligned in the real world. The indicator must first be obtained to access to a particular AR system. It may be inconvenient to have the indicator in reach at all time. Human hand, which is part of the human body may be a solution for this. Besides, hand is also a promising tool for interaction with virtual objects in AR environment. This thesis presents a markerless Augmented Reality system which utilizes outstretched hand for registration of virtual objects in the real environment and enables the users to have three dimensional (3D) interaction with the augmented virtual objects. To employ the hand for registration and interaction in AR, hand postures and gestures that the user perform has to be recognized. Computer vision algorithms are used in detecting the bare hand without the assistance of markers or any other devices such as mechanical devices and magnetic devices. Stereo camera is used to capture video images, so that the depth information of the hand can be computed by using the stereovision approach. Skin color segmentation is employed to segment the hand region from the images and hand features are then extracted from the segmented region, so that the features can be used for the recognition of different hand postures. Instead of fiducial markers, the palm center and fingertips of the outstretched hand are tracked in real time. Incorporating the depth information computed, 3D positions of these hand features are then used to estimate the 6DOF camera pose with respect to the hand,

which in turn allows the virtual objects to be augmented onto the palm accurately. A few simple gestures are mapped for the interaction with the virtual objects. The developed markerless AR system enables the users to inspect the virtual objects from different view angles by simply moving the hand and to select and translate the virtual objects in 3D space with the other hand intuitively. The system achieves an interactive frame rate of 12 fps. The hand posture recognition algorithm developed is able to recognize the hand postures accurately, with overall recognition rate of 97.0%. The accuracy of the camera pose estimation is not very high, but it is sufficient to be used for the registration of virtual objects on the hand.

CHAPTER 1

INTRODUCTION

1.1 Background

Augmented Reality (AR) is an emerging technology which integrates computer generated virtual elements onto the real world in real time. In contrast to Virtual Reality (VR) in which users immersed in virtual environment entirely, AR enhances the users' perception of real world with the virtual elements appear to be coexist in the same space as the real world.

Milgram (Milgram and Kishino, 1994, Milgram et al., 1995) introduced reality-virtuality continuum that defines the merging of real and virtual objects presented in any particular display situation, as shown in Figure 1.1. Real environment and virtual environment are the two extremes on this continuum in which the real environment is defined as an environment that only consists of real objects while the virtual environment only consists of virtual objects. Any environment that falls between these two extremes, where the real and virtual objects coexist in the display, is considered to be a mixed reality environment. Augmented reality and augmented virtuality are both mixed reality environments. The difference between these two environments is that the virtual objects are added onto the real environment in

augmented reality while the real objects are added onto the virtual environment in augmented virtuality.

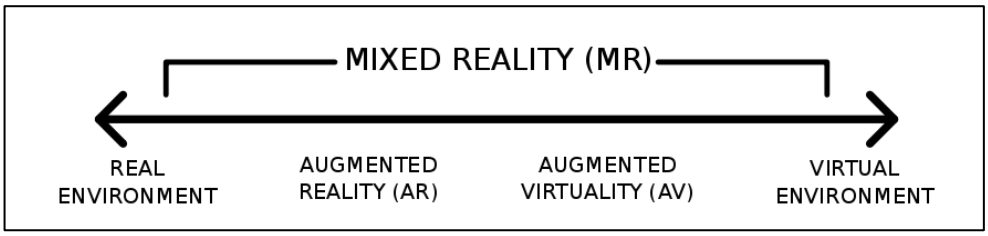

Figure 1.1 Reality-Virtuality Continuum

Azuma (1997) introduced a widely accepted definition for an AR system as a system that combines real and virtual objects in real environment, runs interactively in real time and registers real and virtual objects with each other. This definition allows display technologies other than head-mounted display (HMD) to be employed. Although AR has been applied mostly on sense of sight, the use on all other senses such as hearing, touch and smell is not restricted. Besides, mediated reality where real objects are removed by overlaying virtual objects is also considered AR.

The first AR system which is also the first VR system was developed by Ivan Sutherland and his student, Bob Sproull date back to the 1960s (Sutherland, 1968). Users of this system have to wear an optical see-through HMD that was nicknamed as "The Sword of Damocles" due to its cumbersomeness. One mechanical and one ultrasonic tracker were used to track the position and orientation of the user's head. Based on the information tracked, the virtual objects were rendered on the display of the HMD in real time. With the limitation in processing power of computers at that time, the graphics displayed were simple wireframe drawings.

Since then, researches had been more focused on VR and only until the 1990s, that AR became a significant field of research. The term 'Augmented Reality' was

coined by Caudell and Mizell (1992). AR is a highly multi-disciplinary field of research. The underlying areas which are crucial for an AR system include computer vision, computer graphics, signal processing, user interfaces, human factors, wearable computing, mobile computing, computer networks, distributed computing information access, information visualization and hardware design for new displays. AR is an appealing field as its potential in a number of application domains, such as manufacturing, medical, education, entertainment, military and navigation. The advancement of computing hardware allows rapid development of AR which is more practical for applications.

One of the key components of an AR system is an accurate tracking and registration method to align the real and virtual worlds seamlessly in real time. To register virtual objects accurately in the real world, the viewpoint of the user on the scene is required. Six degrees of freedom (6DOF) that defines the position and orientation of the viewer relative to the scene have to be tracked. Tracking techniques that have been used can be classified into three categories, which are sensor-based, vision-based and hybrid tracking. Sensor-based tracking involves sensors such as mechanical, magnetic, ultrasonic and inertial sensors while vision-based tracking depends on the image processing techniques to compute the camera pose with respect to the scene. Hybrid tracking combines sensor-based and vision-based techniques or combines a few sensors.

Vision-based tracking have been in greater interest for application in AR as it is unencumbered and low cost compared to sensor based tracking. Fiducial markers are often being added to the environment to ease the pose estimation and registration task. Many different types of markers haven been created in the past. For example,

ARToolkit (Kato and Billinghurst, 1999) and ARTag (Fiala, 2005) are two popular marker-based library that have been introduced. ARToolkit uses template markers which are 2D images surrounded by black borders, while ARTag uses ID-based markers, with 6x6 bits pattern confined in 10x10 bits. Despite the fact that marker-based AR is robust and computational simple, the presence of markers is intrusive and unnatural in the real environment. Thus, markerless AR has been explored to eliminate the use of fiducial marker and instead use natural features in the scene for tracking. The tracking for natural features in the scene is certainly a more challenging approach compared to tracking fiducial markers.

Another main aspect for an AR system is the interaction techniques between users and virtual objects in the real environment. An appropriate interaction technique ensures an intuitive manipulation of virtual objects and hence provides a more immersive AR experience to the users. Earlier AR research efforts are more focused on pose tracking and registration of virtual objects. Limited attention has been given on the interaction between users and virtual objects. In the later years, interaction techniques have been growing in interest among the researchers which conceivably reflects the maturity of the field for real world applications (Zhou et al., 2008).

The conventional Human-Computer Interaction (HCI) devices such as keyboard and mouse are not adequate for interaction in AR applications as they only support two dimensional interactions. Multidimensional input devices such as 3D mouse, joystick and PHANToM (Massie and Salisbury, 1994) provide an alternative for interaction in 3D. However, these devices hinder the naturalness in interaction as users are not able to manipulate the virtual objects directly. Tangible user interface,

which allows users to interact with the virtual objects by means of real physical objects have been employed in AR. The main factor contributes to the effectiveness of this interface is the decision on how each motions of the real object associate to the interaction with the virtual objects. The usage of the real objects should be natural and user-friendly to the users without much learning effort. The most natural and promising means of interaction in an AR environment might be using hand gesture as it is a natural way of communication and objects manipulation for human in daily life.

1.2 Motivation

As discussed in the previous section, AR system requires an indicator for the position and orientation where the virtual objects should be rendered in the real environment. To access to a particular AR system, a user has to first obtain the indicator used by the system. For marker-based AR, the indicator is a fiducial marker and it might be any objects for markerless AR which have been defined and learned. It may be sometimes inconvenient to the user to keep the indicator in reach for all time. Thus, an object which is easily accessible by all the users can solve this problem. Therefore, human hand has the potential to serve this purpose well as it is part of the human body.

Besides, hand gesture has been in growing interest as a promising way for interaction with virtual objects in AR environment. It can be used to navigate, select and manipulate the virtual objects, similar to the ways human interact with physical objects. To enable AR system to react to the hand movements, the hand must first be detected and tracked. Earlier researches have exploited the use of mechanical or

magnetic sensor devices such as data gloves in hand detection and tracking. Although this method is found to be more effective and less computationally expensive, several drawbacks arise as the use of data gloves restricts the movement of hand, requires complex calibration and setup procedures. Moreover, mechanical and magnetic devices are mostly expensive. Thus, computer vision approach which is more natural and unencumbered provides a more promising alternative to data gloves. To facilitate the detection and tracking of the hand, some studies have applied markers (Buchmann et al., 2004) or colored gloves (Wang and Popovi, 2009) on the hand. However, an even more particularly desirable interaction approach is by using bare hand without the need of any markers or colored gloves.

1.3 Objectives

The objectives of this research are:

a. To develop an accurate vision-based bare-hand posture recognition algorithm to be used in AR system.
b. To develop a markerless AR system that utilizes hand for registration of virtual objects and enables 3D hand interaction in the AR environment.
c. To evaluate the feasibility of the developed markerless AR system, where both registration and manipulations of the virtual objects are based on hands.

1.4 Scope and Limitation

The key contribution of this thesis is the introduction of a markerless AR system that utilizes outstretched hand for registration of the virtual objects in the real

world and enables users interact with the virtual objects in 3D, by using the other hand. Virtual objects are registered based on the position and orientation of the outstretched hand; and users should be able to select and translate these virtual objects in 3D. The system should meet the challenge to run in interactive frame rate, without requiring the user to wear any additional extra sensor devices or markers.

To enable hands to be used for registration and interaction in the AR environment, hand posture recognition algorithm has to be developed. The algorithm should be able to recognize up to two hands from the images. There are some limitations applied on the recognition, taking the processing time required in consideration. The hand posture recognition algorithm only covers a predefined set of postures, which would be sufficient for registration, selection and translation of virtual objects. The recognition is restricted to controlled background and lighting condition. Only full hand detection is considered, without dealing with the possible occlusion problem of the hands.

1.5 Thesis Organization

This thesis consists of seven chapters and the organization is as follows:

Chapter 1 discusses the background for AR field. The motivation, objectives, scope and limitation for this thesis are introduced.

Chapter 2 reviews related work in the literature. The techniques employed for hand posture and gesture recognition in previous studies are discussed. The reviews focus on vision-based techniques without the use of additional sensors or markers.

Related works that utilize hand for registration of virtual objects and interaction in AR environment are also included.

Chapter 3 presents the overview of the developed AR system. The hardware and software configurations that support the system are discussed. System framework will be introduced in this chapter. Stereo camera calibration and rectification are also included.

Chapter 4 focuses on the techniques for hand posture recognition employed in this system, beginning with the discussion of postures that are supported in this system. The recognition process is divided into three steps, which are hand region segmentation, hand features extraction and hand postures classification for discussion.

Chapter 5 discusses the registration and interaction of virtual objects in AR environment depending on various hand postures and gestures that are recognized.

Chapter 6 presents and discusses the implementation results of the system from the perspective of the performance and accuracy from various aspects.

Chapter 7 concludes the thesis with a brief summary of the work and recommended future research direction in this area.

CHAPTER 2

LITERATURE REVIEW

2.1 Vision-based Hand Posture and Gesture Recognition

Hand gesture is often being referred to both static hand pose and dynamic movement of the hand in many studies. However, some researches adopt two different terms for this two concepts, which are hand posture and hand gesture. In this thesis, the definition for hand posture and hand gesture similar to the one in (Liang and Ouhyoung, 1998) is used to avoid any confusion, i.e. *Hand posture is defined as a static hand pose on a certain time without any movements while hand gesture is a sequence of hand postures result of continuous hand or finger movements over a short time span*.

Human hand has a complex articulated structure consisting of many parts and joints, resulting of a total of 27 degrees of freedom (Ying and Huang, 2001). As a non-rigid and high dimensional object, hand posture and gesture recognition is indeed a very challenging task. The main difficulties in the development of a vision-based hand posture and gesture recognition system as stated in (Erol et al., 2007) include:

- High-dimensional problem : Studies have shown that hand motion involves at least six dimensions. Even taking into account that natural

hand motion involves less than 20 DOF due to independence between fingers, there are still many parameters need to be estimated.

- Self-occlusions : There are a significant amount of possible hand shapes with many self-occlusions depend on the viewpoints, thus it is very difficult to estimate or describe the hand computationally.
- Processing speed : For computer vision approach, a large amount of data have to be processed for each input image. Without efficient algorithm and adequate hardware, the system will not be able to run in real time.
- Uncontrolled environments : The segmentation of an rigid object from even a simple background may not be an easy task, moreover, to segment a non-rigid hand from the environments without restriction such as lighting or background.
- Rapid hand motion : The frame rate supported by the off-the-shelf camera and the algorithms employed is limited. It may not be sufficient to support high speed hand motion.

It is very difficult to develop a system for generic usage. Most developed systems are not able to surmount all the above difficulties and have applied some kinds of restrictions on the users or the environments. Some common restrictions on the environments are to assume the background is static or simple and hand is being the only skin-colored object. There may also be limitation on the hand postures and gestures that are supported by the system to those which are simple and with low DOF. This can ensure minimal self-occlusions of the hand and facilitates a more accurate recognition. Besides, users may be advised to have slower hand motion.

The current approaches for vision-based hand posture and gesture recognition can be categorized into 3D hand model-based approaches and appearance-based approaches (Garg et al., 2009). 3D hand model-based approach depends on a 3D kinematic hand model with considerable DOF. This method matches the input images with the possible 2D appearances projected by the 3D hand model and find the best estimation for the hand parameters. A large database is required to cover all the possible 2D appearances of the hand under different perspective and results in high computational cost for the matching process. Despite the potential of this approach to support for a large variety of hand postures with its rich descriptions, it is hard for this approach to achieve real time performance. Hence, this approach is not suitable for implementation in AR system.

Appearance-based approach models the visual appearance of the hand with its 2D image features and compares the extracted image features from the input image with these parameters. Since this approach deals with 2D image features, the computational cost is lower and thus easier to achieve real time performance. This approach is more suitable for system where only small set of particular hand postures that need to be recognized. The appearance-based approach for hand posture recognition can be generally divided into three steps, which are hand segmentation, hand feature extraction and hand posture classification. Depending on the sequence of hand postures classified over a short time span, hand gesture can be classified. These steps will be discussed in the following subsections.

2.1.1 Hand Region Segmentation

The primary step in hand posture and gesture recognition is to localize the hand

region in the input image and segment it from the background, which often termed as hand segmentation. The quality of hand segmentation result significantly affects the subsequent steps and thus the recognition accuracy.

Skin color segmentation has been widely adopted for hand segmentation as skin color is an important cue for hand and the implementation is simple in computation (Zabulis et al., 2009). As this method is highly invariant to transformations, hand can still be detected even in various hand postures and with self-occlusions. It can work in dynamic background where the camera not necessary to be static. However, this method faces the challenge of removing other objects with similar color such as the face and arm. Besides, color segmentation is very susceptible to lighting conditions. Even for an identical scene, the images captured by different cameras may not have an identical color distribution. The effectiveness of skin color segmentation depends on the skin color model employed. A good skin color model should be insensitive against illumination changes and applicable to humans with different skin tone. Thus, the choice of color space and the method used for modeling the skin color distribution are essential. Several color spaces have been adopted in the previous studies, which include RGB, normalized RGB, YCbCr, HSV, etc. The skin color modeling method can be categorized into three groups, which are explicitly defined skin range, nonparametric skin distribution and parametric skin distribution.

Another method for hand segmentation is background subtraction (Ogihara et al., 2006). Background subtraction generally requires the background to be relatively static and target objects are the only moving objects in the scene. For the traditional background subtraction method, a reference frame which represents the background

scene needs to be obtained. Frame differencing is then being applied to the captured images to determine if a pixel belongs to the foreground moving objects or the background. A pixel is considered as foreground pixel if the difference in intensity between the input image and reference frame is above a predefined threshold. Otherwise, it is regarded as background pixel. However, this method is not suitable for AR systems where the background is not always static as the camera is not stationary to provide greater mobility freedom to the users.

Generally, neither the color segmentation nor background subtraction method alone can provide a robust result of segmentation unless for a static and entirely clean background. As for skin color segmentation, there might be other skin color regions detected in the images. To ensure a better segmentation for the hand region, Alvarez et al. (2010) combine both color segmentation and background subtraction to segment the hand region from the background.

Morphology operations might be applied to the segmented images to eliminate the noises. Hand region is often being assumed to be the largest blob detected in the images. Besides, arm may appear in the segmented images together with the hand region. Seo et al. (2008) removes the arm region from the hand region to ensure a more accurate extraction of hand features in the subsequent step. It is done by first detecting the wrist in the segmented image. Radkowski and Stritzke (2012) apply restriction on the users to wear long sleeve to eliminate the need for step of hand-arm removal.

2.1.2 Hand Feature Extraction

Hand features, which are the useful information of a hand that can be used for identification for its posture, are extracted from the segmented images. These features will be used in the classification process. The choices of features depend on the selection of hand postures and type of interaction that the system aims to support. Some of the important features that are often used are the centroid of segmented hand region (Elmezain et al., 2008), position of palm centre (Raheja et al., 2011), fingertips (Kang et al., 2008) and convexity defect points between the fingers (Wan et al., 2012). Centroid of the hand region may not be a stable point as it varies greatly depends on the existence of fingertips in the images. Location of palm centre is often being referred to as the point with maximum distance to its closest boundary edge. Some of the techniques used to extract the fingertips are curvature-based (Lee and Lee, 2011), template matching (Kerdvibulvech and Saito, 2008) and correlation method (Bilal et al., 2011).

2.1.3 Hand Posture Classification

With the hand features extracted, the detected hand can be classified to its respective posture. Generally the approaches for classification can be divided into two groups, which are rule based and machine learning based (Murthy and Jadon, 2009). Rule-based approaches depend on a set of manually encoded rule for the postures supported. Features extracted from the input images are compared with the encoded rule to obtain the matched hand posture. The accuracy and efficiency of this classification is greatly influenced by the ability of the people who encode the rules to a set that best represents the various hand postures. This approach is more

practical for recognition systems which only involve small set of postures. Machine learning based approaches can provide a more efficient mappings between high dimension feature sets and postures.

2.2 Hand-Based Registration and Interaction

Several researches have worked on markerless augmented reality system which utilizes hand as the tracking pattern for the augmentation of virtual objects on the user's hand. These researches will be discussed in this subsection. In these systems, users are able to inspect the virtual objects conveniently from different viewing angles by moving their bare hand without any additional markers or devices used. Camera pose with respect to the hand is estimated by using the natural features on the hand.

The systems presented in (Lee and Hollerer, 2007), (Lee, 2010) and (Shen et al., 2011) require each of their users to first conduct an offline initialization process to obtain the location of some particular feature points on their hands for the construction of the hand coordinate system. The hand is assumed to be in a plane and thus all the feature points on the hand are considered to be with the same z-coordinate in the hand coordinate system. With the correspondences between the 2D feature points extracted from the input images and the 3D points in the hand coordinate system, 6DOF camera pose relative to the user's hand can be obtained.

The hand features needed varies according to the choice of each particular system. Fingertip is the most common hand feature used to establish the hand coordinate system for camera pose estimation, as in (Lee and Hollerer, 2007), (Lee, 2010) and (Kato and Kato, 2011). In HandyAR system (Lee and Hollerer, 2007), the

hand region is first detected by adaptively learned skin color method. Hand region is assumed to be the largest blob as this system aims for application in wearable computing environment. Fingertips are then detected by curvature-based and ellipse fitting method, and tracked by matching algorithm. The results of this study show that fingertip locations can be used effectively in camera pose estimation without the use of marker. Lee (2010) adopted skin color segmentation with an explicitly defined ranges obtained from an experiment of a set of images. Then, fingertips are detected by curvature-based method and tracked by matching algorithm, similar to the HandyAR system. Kato and Kato (2011) developed a hand-based markerless AR application for smart phones. Due to limited computation power of smart phones, the application adopted a fast fingertip detection algorithm to enable a reasonable real time experience to the users. This application performs segmentation by using online learning of Gaussian statistical color model. Convex hull of the segmented region is computed and its vertices are considered as candidates of fingertips. Non-skin regions within the convex hull are used to identify the fingers. The performance of this method had been compared with the performance of HandyAR and results in the research show that the method used is much faster and accurate than HandyAR.

In (Shen et al., 2011), the four convexity defect points between the fingers are used for camera pose estimation due to the idea that these defect points are relatively more static during hand movements compared to fingertips. Hand segmentation is performed by using color histogram. Fingertips and convexity defect points between the fingers are detected by using curvature-based method.

The hand-based markerless AR system discussed above may not be suitable for use in the event where the systems are constantly accessed by different users. It may

be troublesome for every new user of these systems to have the offline initialization for the hand features. Seo et al. (2008) propose a markerless AR application on mobile devices, where offline initialization of the hand features is not required for the camera pose estimation. Hand region is detected by using generalized statistical color model and with the assumption that hand is the largest blob in the image. The starting point of the forearm, palm direction and the convexity defect point between the thumb and index finger are extracted and used in camera pose estimation. The palm width is determined by a line, which is orthogonal to the palm direction, passes through the convexity defect point and contour of the hand. The palm height is the shortest distance between the line of the palm width and the starting point of the forearm. Four corners of a virtual square are constructed and used for the estimation of the palm pose.

The proposed AR system in this thesis does not require initialization process for the hand features as well. A different method is adopted in this system for camera pose estimation. This system employs stereo camera for image capturing, so that the 3D camera coordinates of the hand features can be computed for camera pose estimation.

Most of the discussed hand-based markerless AR systems are mainly for purpose of inspection of virtual objects on the hand and enable limited interaction between the user and the virtual objects. As for (Seo et al., 2008), the user has to hold the mobile device with one hand and the other hand is used for the registration and interaction with the virtual object. The virtual objects rendered on the palm react according to the opening and closing movements of the hand. The interaction

depends on the fingertips' motions which detected by curvature and ellipse fitting method. This application also enables tactile interaction by means of data glove.

Lee and Hollerer (2007) and Shen et al. (2011) have also implemented the developed hand detection and tracking algorithm for interaction with the virtual objects augmented in the real environment. However, the interaction is in a marker-based AR environment but not in the markerless AR environment developed. Lee and Hollerer (2007) used ARTag to establish a world coordinate system for registration of virtual objects. A user can select a virtual object by moving the hand close to the object, in pixel distance of the image plane. The selected object will then move to the hand and register on it. The object will back to its original position when the user release it. ARToolkitPlus is used by Shen et al. (2011) for the registration of virtual objects. Users are able to use the fingertips to interact with assembly parts in 2D. When the middle point of two fingertips is close to an assemble part, the part will be selected and moved according to the movement of the middle point of the two fingertips.

A noteworthy difference of the work of (Lee, 2010) with the other hand-based markerless AR system discussed is the use of the other hand in the scene as a hand-mouse to interact with the virtual objects. Five hand patterns have been adopted to manipulate the virtual objects, i.e. zoom in, zoom out, rotation about x-axis, y-axis and z-axis. The hand patterns are recognized by using template matching method. This system is most similar to the system developed in this thesis, where one of the user's hand is used for registration of the virtual objects in the real world and the other hand is used to interact with the virtual objects. However, the interaction is restricted to 2D. If there is more than one virtual object rendered in the AR

environment, the user is not able to interact with a particular virtual object, but has to interact with all the rendered virtual objects as one. In the proposed system in this thesis, the interactions is extended to include 3D and enable users to interact with a particular virtual object.

Radkowski and Stritzke (2012) introduce a set of interaction techniques for assembly of the 3D virtual parts in a marker-based AR system, ARToolkit by using hand gesture. Two video cameras are arranged statically in the scene. One of the cameras, Kinect is located opposite to the user and the other webcam is placed next to the user's head. User has to always stand on a fixed position during the interaction. A hand gesture recognition software is implemented in the system. Five gestures, i.e. fist, open hand, closed hand, index finger and waving gesture are supported by the software. The operations allowed in this system include selection, translation, rotation, scaling, change of attributes and virtual assembly. Two modes for interaction have been employed, which are direct mode and precise mode. Direct mode allows the users to manipulate the objects directly where the virtual objects selected will move according to the movement of the hand. Precise mode aims for exact transformation of the virtual parts where a set of choices for the transformation have to be selected by the hand before the objects can be manipulated.

Lee et al. (2008) and Seonho et al. (2013) employ stereovision technique to track hands in 3D and enable 3D hand interaction with the virtual objects in marker-based AR system. Both of these system only support one posture for interaction, which is the 'pointing posture'. By using the center point of hand and the fingertip extracted from the input image, the direction where the finger pointing is computed. Finger ray casting is then used for the collision detection between the user's hand

and the virtual objects. User can select the virtual object if the finger ray collides with the object and move the object if the finger touches the virtual object. Lee et al. (2008) combines hand gesture with speech input for interaction with the virtual objects. The use of speech input in the system enables more interaction with the virtual objects by performing various voice commands. The techniques employed in proposed system in this thesis for 3D hand interaction is similar to these two studies. However, the interaction is based on a hand-based markerless AR system and supports more than one posture.

2.3 Summary

In the chapter, vision-based techniques employed for hand posture and gesture recognition in previous studies are discussed. Related works that utilize hand for registration of virtual objects and interaction in AR environment are reviewed. Only one of the system reviewed utilizes both hands for registration and interaction in AR environment, but this system is restricted the interaction to 2D and users are not able to interact with a particular virtual object if there are more than one virtual object rendered in the AR environment.

CHAPTER 3

SYSTEM OVERVIEW

3.1 System Setup

The proposed system is designed to be used with a personal computer and a video see-through head mounted device, where a pair of stereo camera mounted on it. Stereo camera is used to provide the depth information of an image, by matching the same feature in two images captured by the camera. Head mounted device enables the users to immerse in the AR environment and interact with the virtual objects from a first person perspective. This provides a greater mobility to the users to interact with the virtual objects in AR environment. However, the system is also executable with only normal pair of cameras and other displays.

The programming language used in developing the AR system is C++. OpenCV, an open source computer vision library developed by Intel is used in the implementation of computer vision and image processing algorithms. OpenGL graphics library is used for graphics rendering process.

3.2 System Framework

Figure 3.1 illustrates the system framework. An offline camera calibration and computation of rectification transformation matrix for the stereo camera is first

performed to compute the camera intrinsic parameters and the geometrical relationship between the two cameras. When the AR system executes, stereo images will be captured, and then be corrected for distortion and rectified. Hand posture recognition will then be performed on the left and right images simultaneously to detect the hand postures occurred in the images. If outstretched hand is detected from the images, camera pose relative to the hand is estimated and used in the augmentation of the virtual objects on the hand. If other valid postures are detected, users are able to select and translate the virtual object with different gestures. Detailed explanations of each step are included in the next section, Chapter 4 and Chapter 5.

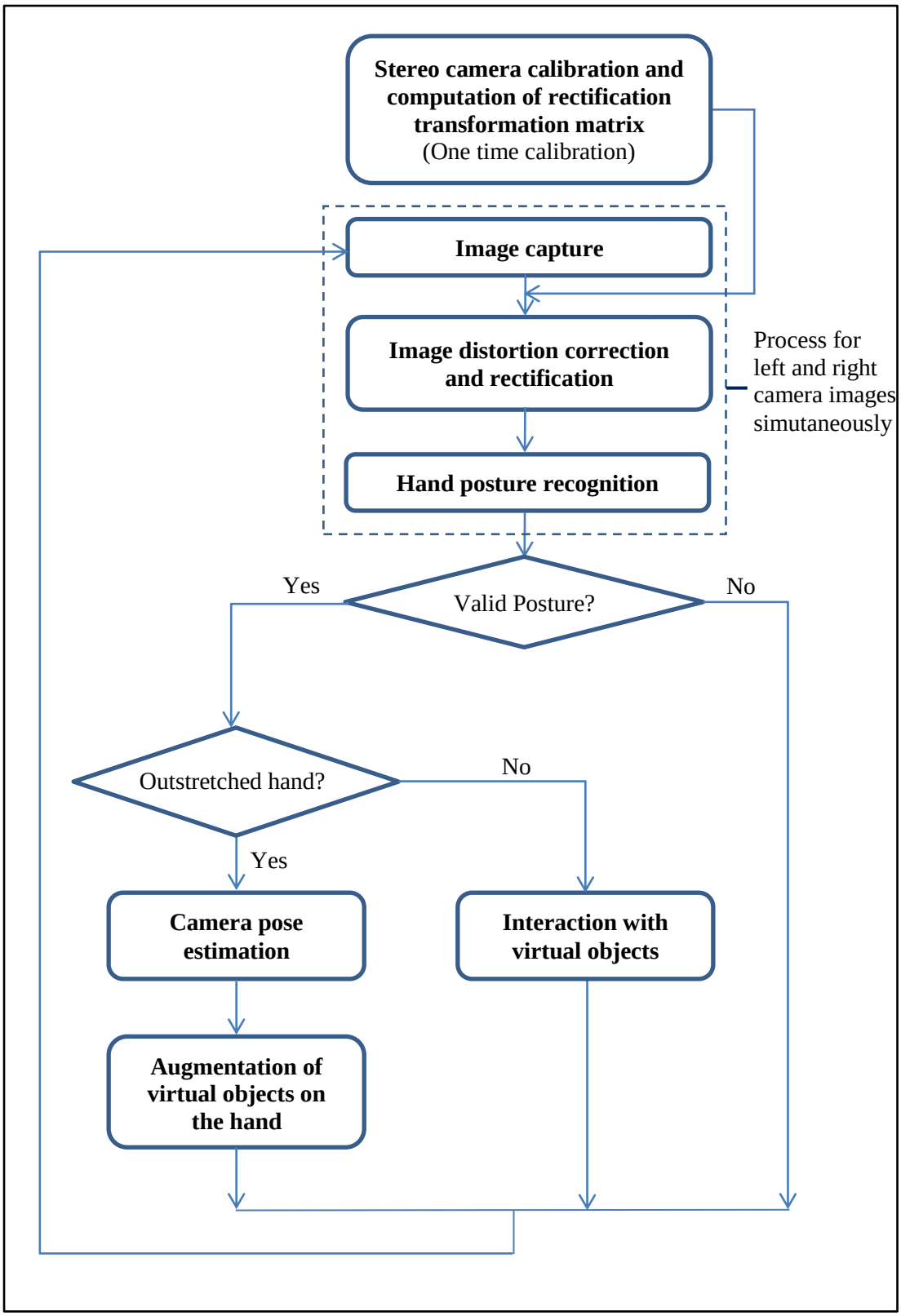

Figure 3.1 System framework

3.3 Stereo Camera Calibration and Rectification

Stereo camera calibration and rectification are essential for the reconstruction of 3D world from the captured stereo images, which makes interaction in 3D world possible. Camera model describes the projective mapping between 3D world and 2D image plane of a camera. In this system, simple pinhole camera model is used.

Stereo camera calibration aims to compute camera intrinsic parameters and the geometrical relationship between the two cameras. Intrinsic parameters specify camera's internal characteristics such as its focal length, principal point and lens distortion coefficients. With the focal length and principal point obtained, the perspective transformation to project the 3D world into the image plane can be determined.

The lens distortion coefficients are used to remove the distortion occurred in the images. The use of lens in the camera often results in distortions, especially on non-professional cameras. The two main types of lens distortions are radial distortion and tangential distortion. Radial distortion occurs due to the shape of the lens. There are two types of radial distortion, which are barrel distortion and pincushion distortion (as illustrated in Figure 3.2). For barrel distortion, image magnification decreases with distance from the optical axis. Thus, straight lines are visibly curved inwards, especially for pixels near the edge of the image. Pincushion distortion happens when image magnification increases with distance from the optical axis and results in straight lines appear to be pulled upwards in the corner. Tangential distortion occurs due to manufacturing defects, where the lens is not being aligned exactly parallel to the image plane.