DUAL-SENSOR APPROACHES FOR REAL-TIME

ROBUST HAND GESTURE RECOGNITION

by

Kui Liu

APPROVED BY SUPERVISORY COMMITTEE:

_____

Dr. Nasser Kehtarnavaz, Chair

_____

Dr. Roozbeh Jafari

_____

Dr. Issa M. S. Panahi

_____

Dr. P. K. Rajasekaran

To my beloved family

DUAL-SENSOR APPROACHES FOR REAL-TIME ROBUST HAND GESTURE RECOGNITION

by

Kui Liu, BS, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

May 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Nasser Kehtarnavaz, for his invaluable guidance and support for my research. He was so kind to have granted me the opportunity to work on this research project and was always willing to discuss the difficulties during the entire course of the project. He has been an excellent mentor and has provided valuable guidance whenever I approached him with problems related to the project. I also am extremely grateful for his constant support and many suggestions during this dissertation work.

I am also very thankful to the other members of my committees, Dr. Roozbeh Jafari, Dr. Issa Panahi and Dr. P. Rajasekaran. The discussions I had with them during my Dissertation Proposal were valuable to improve this dissertation work.

March 2015

PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper or papers submitted for publication. The dissertation must still conform to all other requirements explained in the "Guide for the Preparation of Master's Theses and Doctoral Dissertations at The University of Texas at Dallas." It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student's contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

DUAL SENSOR APPROACHES FOR REAL-TIME

ROBUST HAND GESTURE RECOGNITION

Publication No. _____

Kui Liu, PhD
The University of Texas at Dallas, 2015

Supervising Professor: Nasser Kehtarnavaz

The use of hand gesture recognition has been steadily growing in various human-computer interaction applications. Under realistic operating conditions, it has been shown that hand gesture recognition systems exhibit recognition rate limitations when using a single sensor. Two dual-sensor approaches have thus been developed in this dissertation in order to improve the performance of hand gesture recognition under realistic operating conditions. The first approach involves the use of image pairs from a stereo camera setup by merging the image information from the left and right camera, while the second approach involves the use of a Kinect depth camera and an inertial sensor by fusing differing modality data within the framework of a hidden Markov model. The emphasis of this dissertation has been on system building and practical deployment. More specifically, the major contributions of the dissertation are: (a) improvement of hand gestures recognition rates when using a pair of images from a stereo camera compared to when using a single image by fusing the information from the left and right images in a

complementary manner, and (b) improvement of hand gestures recognition rates when using a dual-modality sensor setup consisting of a Kinect depth camera and an inertial body sensor compared to the situations when each sensor is used individually on its own. Experimental results obtained indicate that the developed approaches generate higher recognition rates in different backgrounds and lighting conditions compared to the situations when an individual sensor is used. Both approaches are designed such that the entire recognition system runs in real-time on PC platform.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Human-computer interface (HCI) is an active research area in computer vision that emerged in the early 1980s. It involves utilizing both human and machine to achieve a specific task. Examples of the human-computer interface technology include face recognition, speech recognition, eye tracking, gesture recognition and wearable body sensing. Normally, a single sensor or a single modality sensor is deployed for HCI. However, the use of a single sensor or a single modality sensor has limitations when operating under realistic conditions. This dissertation aims at exploring the use of more than one sensor in order to achieve a more robust hand gesture recognition. More specifically, the hand gesture recognition problem is examined by considering a pair of the same modality sensors and a pair of differing modality sensors. It is hypothesized that fusion of information from dual sensors would improve the recognition outcome under realistic operating conditions. One fusion scenario involves using two sensors of the same modality or a pair of stereo cameras, and the other fusion scenario involves using two sensors of differing modalities or a depth camera and an inertial sensor. Real-time application has been the thrust of this dissertation, that is placing emphasis on system building or practical deployment rather than pure theoretical development.

Experiments reported in the literature for hand gesture recognition are mostly conducted under controlled lighting and backgrounds when using a camera. However, in practice, image information captured under different lighting conditions and in various backgrounds drastically

changes the recognition outcome. Similarly, when using a wearable body sensor for hand gesture recognition, sensor jitters and drifts occur and as a result many false alarms get generated. In this dissertation, it is hypothesized that the simultaneous utilization of dual sensors of the same modality or differing modalities would increase the recognition system robustness by using the information from dual sensors in a complementary way. Naturally, there are challenges when fusing data from two sensors. When fusion is done across two sensors of the same modality, the data collected might contain redundancy. The data fusion approach should be designed in such a way that redundant information are identified and not used. When fusion is done across two differing modality sensors, data sample correspondences need to be established. For example, up-sampling and down-sampling may be necessary to make sure data frequencies match.

1.1 DIFFERENT TYPES OF FUSION

Fusion of information from two sensors can be done in different ways. In general, one may perform three types of fusion:

1) Data-level fusion
2) Feature-level fusion
3) Decision-level fusion

As shown in Figure 1.1, data-level fusion occurs at the data level where the incoming raw data from different sensors are combined. This type of fusion can be applied in cases when data are of the same type. In other words, when sensors of the same modality are used. Data-level fusion combines raw data from several sensors of the same modality for a classifier. Another type of fusion is feature-level fusion. Feature-level fusion involves carrying out fusion of features after features are extracted from raw data. This type of fusion requires carrying out the

processes of synchronization, noise suppression, down-sampling and up-sampling. The fusion type with the least computational complexity is decision-level fusion. It involves fusing the decisions made by individual classifiers or decision makers. In this dissertation, the feature level fusion is considered.

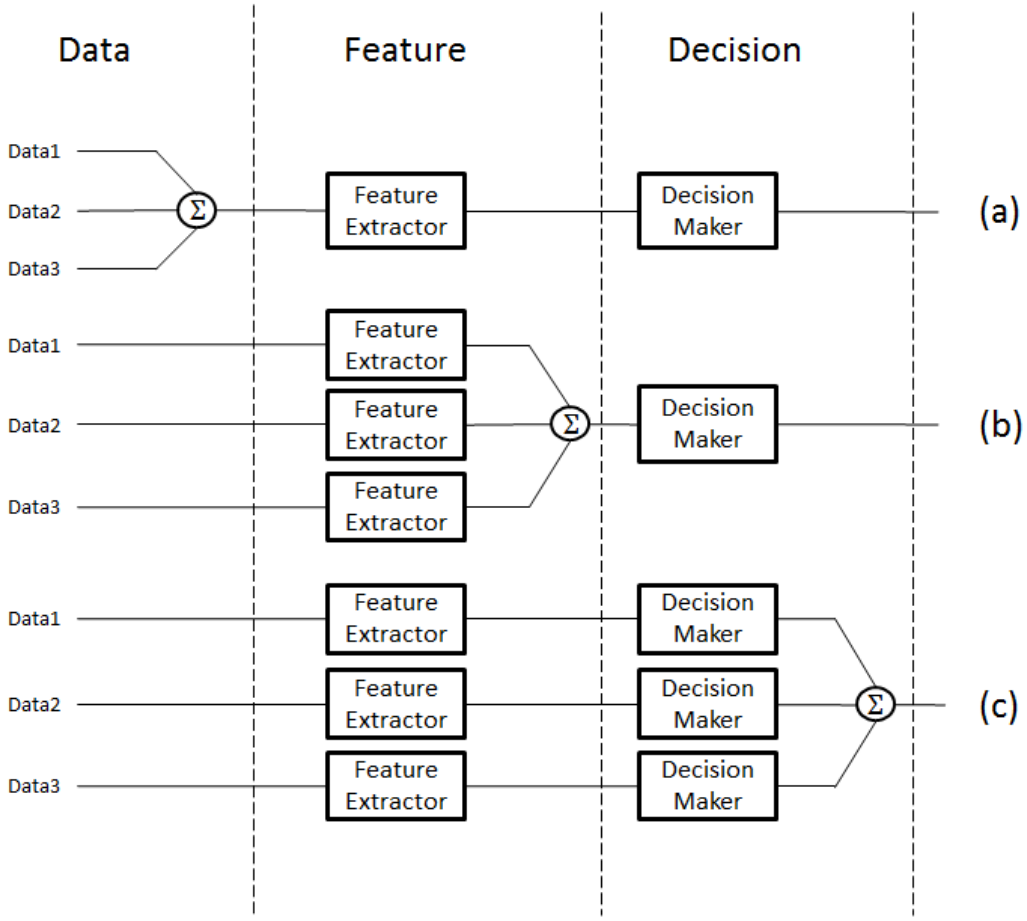

Figure 1.1. Three different types of fusion: (a) data-level fusion, (b) feature-level fusion, (c) decision-level fusion.

1.2 HAND GESTURE RECOGNITION METHODS

The goal in hand gesture recognition to interpret human hand gestures via computer algorithms. For instance, sign language and hand motions are used to provide human communication. Both non-vision and vision-based approaches have been used to achieve hand gesture recognition. An example of a non-vision approach was reported in [1] where finger bending was detected by a pair of wired gloves. In general, vision-based approaches are more natural as they require no hand fitting devices. In all the vision-based hand gesture recognition methods, the task to accomplish is the same. That is, given a sequence of hand images, measure the similarity between the image sequence and a hand model. Vision-based approaches can be divided into active and passive sensing. Active sensing approaches have proven successful for hand gesture recognition, in particular through the use of Kinect [2] [3] and time-of-flight depth cameras [4].

Many passive vision-based hand gesture recognition techniques have been introduced in the literature, e.g. [5-7], where images from a single camera are used to achieve hand gesture recognition. Hand gestures can be classified into two categories: static and motional gestures. A recognition technique for static gestures was reported in [8], where features derived from elastic graph matching was used to identify hand postures in complex backgrounds leading to a recognition rate of 85%. In [9], a learning approach based on a disjunctive normal form was used leading to a recognition rate of 93%. This approach involved the use of normalized hand moments and compactness. In [10], finger spell recognition was achieved at the processing rate of 125ms per image frame using the CamShift algorithm. In [11], principal component analysis was used for hand gesture recognition.

As far as motional hand gesture recognition is concerned, the following three major approaches have been utilized: optical flow, model-based and HMM. In [5], a hand gesture model was devised using an Adaboost classifier and Haar features together with a Kalman predictor to cope with false detection. In [7], a model-based tracking of hand gestures was considered. The use of HMM for hand gesture recognition was discussed in [12].

1.3 STEREO FUSION

Stereo camera is a type of camera that has two optical sensors or cameras as shown in Figure 1.2. In many cases, the distance between the optical centers of a typical stereo camera is comparable to the distance between human eyes. Stereo cameras are often used to create a sense of depth perception.

Figure 1.2. Stereo cameras

The first approach developed in this dissertation involves performing hand gesture recognition in real-time based on a pair of stereo images by fusing the information from the left and the right camera of an inexpensive stereo webcam, such as the one in [13]. This method is different from the traditional high computational complexity method of depth image generation from the stereo camera. The attempt made here is to increase the robustness of hand detection

and hence the robustness of hand gesture recognition by using a pair of low-resolution stereo images, instead of images taken from a single camera. The challenge in this attempt is to increase robustness in a computationally efficient manner so that a real-time throughput is achieved. The developed approach establishes a balance between robustness and computational complexity. On one hand, the solution is designed to be robust to different backgrounds and lighting conditions. On the other hand, it is designed to incorporate time-efficient and relatively simple functions to achieve a real-time throughput. The developed approach combines or merges the information from the left and right images of a stereo camera in order to increase the robustness of hand detection while meeting the real-time constraint.

Figure 1.3 represents the stereo disparity geometry where a left and aright image are captured by a pair of cameras. In this figure, *pl* and *pr* denote the projected points of a scene point into the stereo images. The disparity of the scene point in the left and right images translates into a horizontal displacement of the projected points that are located on so called the epipolar line:

$$d_{lr} = X_l - X_r. \tag{1}$$

where X_l and X_r represent the horizontal coordinates of the projected pixels in the left and right images, respectively. Given that the vertical disparity can be rectified, the depth of the scene point can be computed from the following equation:

$$z = b \times f / d_{lr} \tag{2}$$

where b denotes the distance between the camera lens centers (baseline) and f the focal length of the cameras which are normally identical. It is seen that depth is inversely proportional to

disparity in Equation (2). In other words, an object closer to the stereo camera generates higher disparity.

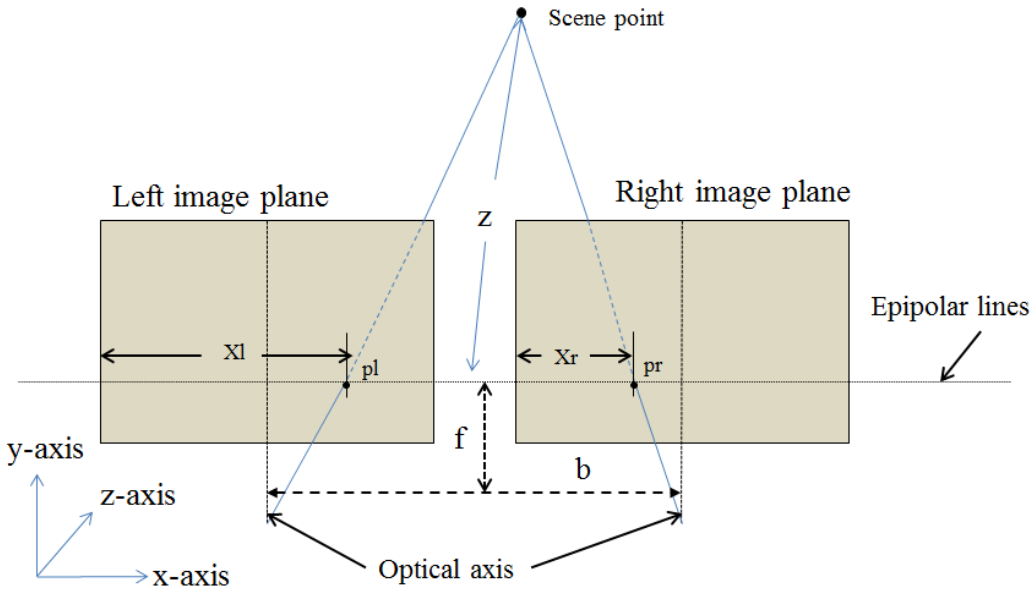

Figure 1.3. Stereo disparity geometry

1.4 DUAL-MODALITY FUSION

The aim of dual-modality fusion is to gain complementary information from differing modality sensors where data from a single modality cannot provide sufficient recognition accuracy.

1.4.1Kinect

Microsoft's Kinect is a low-cost RGB-Depth sensor introduced by Microsoft for human-computer interface applications.. A similar sensor is Xtion pro introduced by ASUS.

Kinect is the first large-scale, commercial release of a depth sensor device and its application have extensively grown in many human computer interaction applications. For

instance, this device is used as a game controller by analyzing 3D data using open-source algorithms for feature selection, scene analyzing, motion detection, skeleton tracking, face recognition and gesture recognition. It deploys an infrared projector and an infrared sensor to obtain the depth information. Two software packages are publically available for this sensor (OpenNi/NITE and Kinect SDK) that allow performing feature selection, motion tracking, gesture and movement recognition. These software packages can be used to track the body joints as illustrated in Figure 1.4. They allow Kinect to recognize people and track their actions. Using the infrared camera, it can recognize up to six objects in the field of view of the sensor. Two objects out of six can be tracked relatively accurately. The joints of users and their movements can be tracked in space and time. The skeleton tracking capability enables recognizing objects standing or sitting. To be recognized, objects need to be in front of Kinect, making sure that the head and upper body are visible; it requires no specific calibration.

The introduction of Kinect has led to successful recognition in many applications including video games, virtual reality and gesture recognition. Not only of its low cost, Kinect is selected as one of the sensing devices here because of its ability to cope with 3D gestures in real-time. Figure 1.5 gives an example of a hand gesture depth image generated by Kinect. In the near range mode, Kinect can detect objects at distances between 0.4 and 3.0 meters.

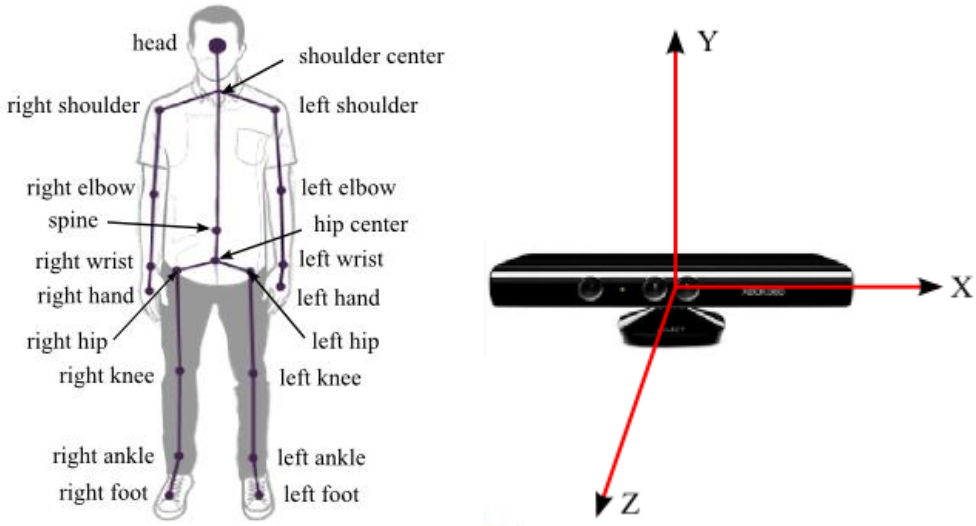

Figure 1.4. Skeleton joints of a human body and Kinect world coordinates

Figure 1.5. Depth map generated by Kinect depth camera

1.4.2 Inertial Sensor

Expensive inertial sensors such as strategic and tactical sensors which provide low bias are often utilized in defense and commercial aviation. Low-cost consumer inertial sensors with typical bias (>30deg/h) are often used in mobile and video game devices such as Nintendo Wii. Inertial body sensors, which produce acceleration and angular signals, are now well-developed

allowing various types of motion to be monitored. However, inertial body sensors normally produce drifts in dead reckoning.

Figure 1.6 shows a 9-axis wireless body sensor having a size of 1”x1.5” that was designed and built in the ESSP Laboratory at the University of Texas at Dallas. It consists of (i) an InvenSense 9-axis MEMS sensor MPU9150 which captures 3-axis acceleration, 3-axis angular velocity and 3-axis magnetic strength data, (ii) a Texas Instruments 16-bit low power microcontroller MSP430 which provides data control, (iii) a dual mode Bluetooth low energy unit which streams data wirelessly to a laptop/PC, and (iv) a serial interface between MSP430 and MPU9150 enabling control commands from the microcontroller to the MEMS sensor and data transmission from the MEMS sensor to the microcontroller. For the magnetometer to provide an accurate reference, a controlled magnetic field without any distortion is required. Thus, here the 6-axis data consisting of 3-axis accelerometer and 3-axis gyroscope are used noting that a controlled magnetic field is not normally available.

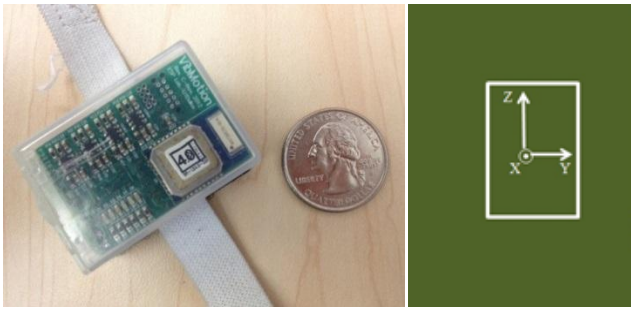

Figure 1.6. Wireless inertial sensor and its world frame

1.4.3 Fusion of Kinect and inertial sensor

For hand gesture recognition and movement monitoring, wireless inertial body sensors have been used due to their low cost and ease of use. The developed real-time hand gesture recognition system is based on the simultaneous utilization of a low-cost inertial body sensor and

a low-cost Kinect camera in a collaborative way in order to detect and recognize a set of hand gestures appearing in the Microsoft MSR dataset [14]. The utilization of both of these two sensing devices (Kinect camera and inertial motion unit (IMU) or sensor)) at the same time is expected to provide system robustness for practical deployment. The fusion or merging of data from these differing modality sensors has not been previously explored for this purpose.

The literature includes a large collection of works where either vision sensors or inertial body sensors have been used for measurement or recognition of human body movements spanning various applications including healthcare rehabilitation and consumer electronics entertainment, e.g. [15-22]. Each of the above two sensors has been used individually for body movement measurements and recognition. However, each sensor has its own limitations when operating under realistic conditions. The major contribution of this work is the development of a general purpose fusion framework to increase the robustness of measurement or recognition by utilizing the information from two differing modality sensors at the same time and in real-time. The above two sensors are deployed in such a way that they act in a complementary manner by compensating for erroneous data that may get captured by each sensor individually. The focus of this work is on hand gesture recognition. However, it should be noted that the approach is general purpose in the sense that the same idea can be applied and extended to other types of human body movements. More specifically, this work involves the fusion of data from a cost-effective inertial body sensor and a cost-effective depth sensor in order to achieve more robust hand gesture recognition compared to the situations when these sensors are used individually.

As far as vision sensors are concerned, two major matching techniques have been deployed for hand gesture recognition. These techniques are Dynamic Time Warping (DTW)

[23] and Elastic Matching (EM) [24]. Statistical modeling techniques such as particle filtering [25] [26], and Hidden Markov Model (HMM) [27] have also been utilized for hand gesture recognition. The application of depth sensors, in particular Kinect [28], has been steadily growing for body movement measurements and recognition. Several studies utilizing the depth sensor Kinect have been reported in the literature for hand gesture recognition. For example, in [2], depth images captured by Kinect were used to achieve recognition of American Sign Language (ASL). In [3], both depth and color information captured by Kinect were used to achieve hand detection and gesture recognition. In [4], a Kinect-based rehabilitation system was developed to assist patients in recovering their muscle atrophy and cerebral palsy. In [29], a HMM was trained to identify the dynamic gesture trajectory of seven gestures using the Kinect sensor.

As far as inertial body sensors are concerned, many body measurement and recognition systems involving such sensors have been presented in the literature. For example, a human motion capture system using wireless inertial sensors was presented in [18]. In [19], wireless body motion sensors were used to recognize the activity and position of the upper trunk and lower extremities. In [20], a customizable wearable body sensor system was introduced. In [21], a SVM classifier was used as part of a body sensor network to estimate the severity of Parkinsonian symptoms. In [22], Kalman filtering as part of a body sensor network was used to obtain dynamic orientations and positions of body limbs.

The simultaneous utilization of both inertial body sensor and depth sensor in real-time has been fairly limited in the literature. In [30], an angle estimation approach involving both an inertial sensor and a Kinect senor was discussed where Kalman filtering was applied to correct or

calibrate for the data drifting of the inertial sensor. The fusion approach done here differs from all the previous works in the sense that both inertial and depth sensor data are used at the same time and together as the input to a probabilistic classifier in order to increase the robustness of recognition. Another attribute of this approach is that the computational complexity is kept low so that its real-time implementation is made possible. Furthermore, both of the sensors deployed are cost-effective which eases their joint utilization in various applications. The developed approach uses HMM classification as this classifier has been proven effective in various recognition applications due to its probabilistic framework. It is also worth stating that this is the first time HMM is used to fuse the signals from a Kinect depth camera and an inertial sensor.

1.5 TEST DATABASES

For the stereo fusion approach, 50 cases of a movement were considered under various background and lighting conditions consisting of florescent, lowlight, sunlight, incandescent lighting. The input images were captured with the stereo webcam Novo Minoru, which is an inexpensive stereo webcam generating low resolution images of size 640*480. Additional stereo images were examined using a Fuji stereo digital camera. The CCD sensors of this camera support 2-10 MP resolution.

There are a number of datasets for testing hand posture or gesture recognition [31], [32] and [33]. However, most either include 2D hand gestures or two-handed gestures. For the dual-modality fusion approach, two gesture sets were considered. One gesture set considered was the single hand gestures in the Microsoft Action Dataset [14] and the other gesture was the $1 Gesture Recognizer Dataset [34]. There are 5 single hand gestures in the Microsoft Action

Dataset and 15 single hand gestures in the $1 Gesture Recognizer Dataset. The hand gestures in the $1 Gesture Recognizer Dataset are used to manage and navigate an Opera Web Browser [35].

1.6 CHAPTER ORGANIZATION

The rest of the dissertation chapters are organized as follows. In Chapter 2, the existing techniques are discussed in detail. The developed fusion approaches are then covered in Chapter 3. In Chapter 4, a comparison between the stereo fusion approach and the dual-modality fusion approach is presented. Finally, the dissertation is concluded in Chapter 5.

CHAPTER 2

EXISTING TECHNIQUES USED IN HAND GESTURE RECOGNITION

Many techniques have been developed for data fusion within two main processes of object tracking and signal matching. Object tracking involves online-calibration, detecting moving objects of interest, and segmentation. Signal matching involves correlating a well-trained signal or template signal with a test signal to detect the presence of the template in the test signal.

2.1 GAUSSIAN MIXTURE MODEL

Gaussian Mixture Model (GMM) is a well-established method for density estimation. It is commonly used as a probabilistic model for representing distributions of features such as the sound intensity and frequency of specific words in speech recognition and the skin color cluster in face recognition. Normally, GMM parameters are estimated by using the iterative Expectation-Maximization (EM) algorithm based on a large training dataset or by using the Maximum A Posteriori (MAP) algorithm [36] from a well-trained prior model. In [37], GMM was used to solve image inverse problems via piecewise linear estimations. In [38], GMM was applied to recover the spatial images contaminated by noise. In [39], an advanced color correction method RACE [40] was utilized to correct the skin-color and GMM was employed to describe the hand colors.

2.2 MEANSHIFT AND CAMSHIFT

Meanshift algorithm is a nonparametric iterative algorithm which uses a generalized kernel. Originally, it was used as a cluster segmentation method. It requires no prior knowledge of the number of clusters.

Recently, Meanshift [41] [42] and its variation Camshift [43] [44] are mostly used to do visual tracking. The window size used to do tracking is the only parameter which needs to be estimated in the Meanshift algorithm. In Camshift, the window size is adaptive with the updating of the convergence of Meanshift. The major disadvantage of Meanshift and Camshift is that they are computationally expensive and do not scale well with the growing dimensionality of the feature space.

2.3 CONVEX HULL

The convex hull of a point set is the smallest convex space which contains the points. For a finite 2D point set, the convex hull is the smallest convex polygon containing all the points. Computationally efficient algorithms such as the Quick Hull algorithm [45] exist for computing convex hulls. The worst-case complexity of this algorithm for a point-set containing n points is $O(n\log n)$. Therefore, the computational efficiency aspect of the convex hull makes it particularly suitable for real-time recognition tasks.

2.4 DYNAMIC TIME WARPING

The Dynamic Time Warping (DTW) algorithm has been successfully used to evaluate the similarity between two given temporal sequences which are varied in time and speed [46]. Initially, DTW was used to compare a prototypical model (template) of words in automatic

speech recognition [47] [48]. Later. new and more computationally efficient DTW algorithms have been developed [49]. This algorithm has been successfully used to cope with speed variations of time sequences [50] [51]. In spite of its $O(n^2)$ complexity, it is extensively used in time series matching problems.

In DTW methods, different distances are used. In particular, Manhattan distance [52], Euclidean distance [49] and Lp Norms [53] are the most common. For all the three distances, the Euclidean distance method provides linear computational complexity and also the implementation is the most straightforward without any parameter setting. All of the distances, however, are sensitive to temporal misalignments. Time shifted or expanded signals are difficult to be recognized.

2.5 HIDDEN MARKOV MODEL

Hidden Markov Model (HMM) is a widely used statistical model originally utilized for speech recognition. In [54], the HMM toolkit HTK was applied to the features extracted from a filterbank for speech recognition. In [55], a recognition system using HMM and a simple pattern matching was utilized to predict characters based on their online writing information. In [56], a data fusion framework was presented which combined HMM and SVM to recognize gestures and postures. Features of the gestures were extracted using statistical properties. The decisions of HMM and SVM were then integrated to improve the recognition rate. In [57], skin color segmentation was performed in the YCrCb space, followed by Kalman filtering and HMM to recognize Malaysian Sign Language (MSL).

2.6 SUMMARY

In this chapter, the major existing techniques previously deploted for hand gesture recognition were presented, including Gaussian mixture model, Meanshift and Camshift tracking, Convex Hull, Dynamic Time Warping and Hidden Markov Model. In the subsequent chapters, these techniques will be revisited within the context of the developed fusion approaches.

CHAPTER 3

REAL-TIME HAND GESTURE RECOGNITION USING DUAL SENSORS OF THE SAME AND DIFFERING MODALITIES

3.1 INTRODUCTION

Hand gesture recognition enables humans to use a most versatile instrument – their hands – in a natural and effective way to perform HCI. Hand gesture recognition by using passive and non-intrusive sensors is often preferred over those using intrusive sensors.

The fusion strategy utilized in this dissertation is based on using the same or different modality sensors that are low-cost. It is to be noted that the introduced approaches to hand gesture recognition here are general purpose in the sense that they can be applied to other human body movements applications. Portions of this chapter have been previously published in reference [23], [88], [89] and [90].

3.2 PREVIOUS WORKS ON STEREO FUSION

The use of stereo images for real-time passive vision-based hand gesture recognition has been previously addressed in the literature. In [13] and [14], a stereo camera with dedicated hardware was utilized to generate depth maps for hand gesture recognition; however, no real-time processing rates were reported in these references. In [58], multiple sets of stereo camera coordinate systems was calibrated and converted into the same world coordinate system to

generate a point cloud of the face and the hand. Skin color detection was used to track the hand. HMM was utilized to achieve a recognition rate of 89.6% for American Sign Language (ASL) by tracking the center of the hand. In [59], a real-time hand gesture system based on stereo cameras was presented. A depth map was generated to detect the hand by using the convex hull technique, then a thinning method was utilized to recognize hand gestures based on hand feature points, their angles and distances. A recognition rate of 83% was achieved based on five types of hand gestures. In [60], a multi-camera system was utilized to detect face and hand gestures. Two fixed cameras were used to predict the hand and face position, while the other two moving cameras were used to track face and hand targets based on the skin color technique. Template matching was used in [61] to recognize hand gestures.

3.3 STEREO FUSION

Two types of hand gestures have been considered in the stereo fusion approach: directional hand movement and finger number spelling. The developed recognition system consists of four main components: online color calibration of hand color, color-based hand detection, hand tracking, and finally hand gesture recognition.

3.3.1 Online color calibration

The goal of the online color calibration component is to adapt subsequent color processing to the color characteristic of the light source under which images are captured. This technique has been previously used quite successfully for face detection in [62] in order to cope with unknown color characteristics of various light sources encountered in practice. The calibration is done at the beginning and only once when the system is turned on. It involves building a GMM model in the CrCb color space to represent the color characteristics of the hand

being captured in an online or on-the-fly manner. The calibration is performed easily by the user simply placing his or her hand in a box displayed at the image center, see Figure 3.1. Representative skin color pixels are collected within this box using a two-cluster k-means clustering algorithm separating skin pixels from non-skin pixels. A GMM model is then trained and used for a region growing hand color segmentation within a region-of-interest specified by a tracking module mentioned next. More details of the online color calibration can be found in [62].

(a) (b)

Figure 3.1. Online color calibration: (a) left camera calibration box, (b) right camera calibration box

3.3.2 Hand detection

There are two main steps for hand detection involving stereo fusion which includes hand tracking and improving robustness based on stereo images.

3.3.2A Hand tracking

The existing tracking methods including optical flow, either sparse [63] or dense [64], and Kalman filtering pose challenges as far as the real-time aspect is concerned due to their

computational complexity. To have a computationally efficient tracking, the CamShift algorithm [43] is adopted here. In this algorithm, the hue component of color is used for tracking. Figure 3.2 shows a sample hue histogram associated with a hand. The histogram within a window is used as the hand tracking feature together with a searching window. The center of the window is used as the seed point for the so called flood fill region growing operation [65] to achieve segmentation in a computationally efficient manner. The CamShift algorithm works similar to the MeanShift algorithm but it also copes with dynamically changing distributions by readjusting the search window size.

Figure 3.2. Sample hue histogram used for CamShift hand tracking

The segmented areas from the left and right images are merged by aligning the left and right images as was previously reported in [66]. The merged area is then used for hand contour extraction. The flow chart of all the components involved in the developed approach appears in Figure 3.3.

3.3.2B Increasing robustness via stereo images

The following rules are introduced to merge the information from the left and right cameras leading to more robust hand detection as compared to using a single camera image. Let x_{i+1} denote the current hand mask and x_i the previous hand mask. The superscripts l and r indicate the left and right camera label for the masks. Let S represent the mask area and δ a percentage parameter reflecting the mask area difference between the frames. The experimentations done have revealed that a δ value in the range 25%-30% can cope with the variability in hand motions made by various subjects. Due to the continuity of motion, it is not physically possible to have a large mask area difference between the frames either left to right or current to previous. Even when the hand is approaching the camera, the mask area is expected to grow consistently. Only when the current mask does not exhibit a large difference from the previous one in both the left and right images, the masks get merged. For instance, first the mask areas between a current left frame $S(\mathrm{x}_{i+1}^{l})$ and a current right frame $S(\mathrm{x}_{i+1}^{r})$ are compared. If there exists relatively little difference between them as per Equation (3), the change is considered to be consistent. Next, the change in the mask areas between a previous and a current frame is examined. If both of the left and right area changes, that is $|S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i}^{l})|$ and $|S(\mathrm{x}_{i+1}^{r}) - S(\mathrm{x}_{i}^{r})|$, show a consistent change between a current and a previous image as per Equation (4), the current left and right hand masks are merged and get updated as per Equation (5). If one of the areas leads to an inconsistent change as per Equation (6) or (8), the update process in Equations (7) and (9) is done. If there exists a large difference between a current left or a current right frame and the image side which does not change consistently as per Equations (11) and (13), it is not

used for the next time frame and only the consistent image side is used as per Equations (12) and (14). Otherwise, no update is done for the next time frame as per Equation (15).

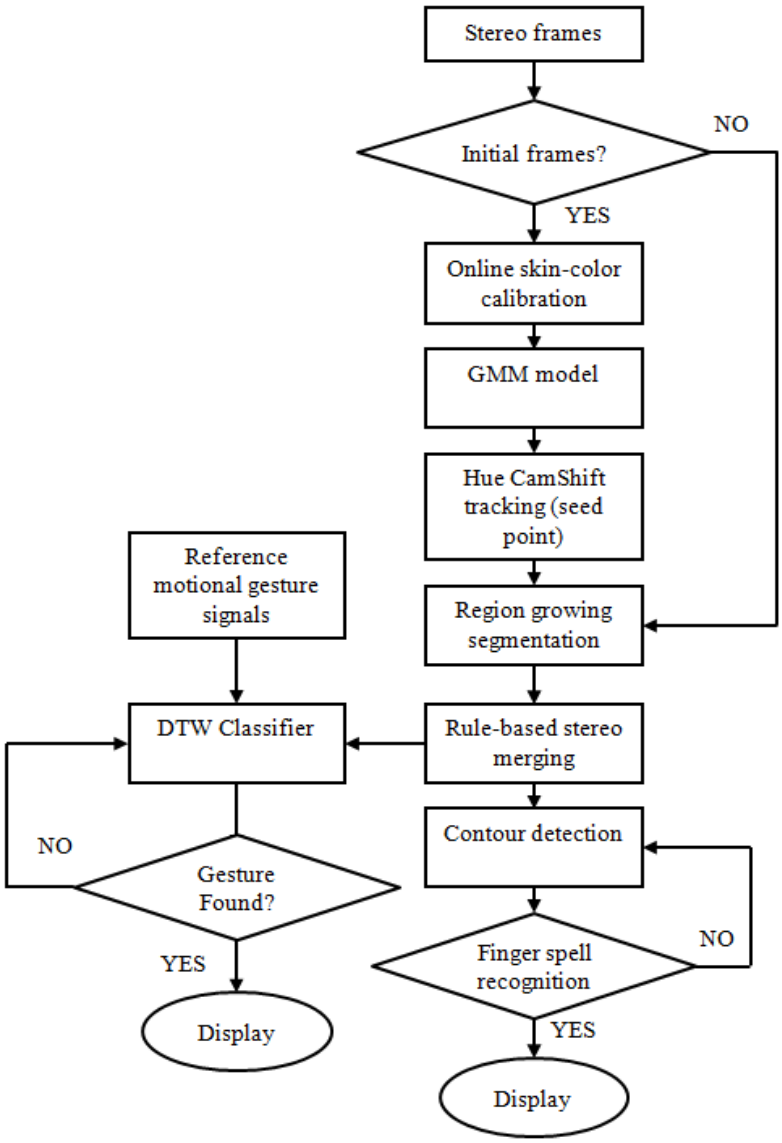

Figure 3.3. Flowchart of the introduced real-time solution using stereo images

$$\text{if } |S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i+1}^{r})| < \delta * S(\mathrm{x}_{i+1}^{l}) \tag{3}$$

$$\text{if } |S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i}^{l})| < \delta * S(\mathrm{x}_{i+1}^{l}) \text{ and } |S(\mathrm{x}_{i+1}^{r}) - S(\mathrm{x}_{i}^{r})| < \delta * S(\mathrm{x}_{i+1}^{r}) \quad (4)$$

$$\mathrm{x}_{i+1} = \mathrm{x}_{i+1}^{l} + \mathrm{x}_{i+1}^{r} \quad (5)$$

$$\text{elseif } |S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i}^{l})| < \delta * S(\mathrm{x}_{i+1}^{l}) \text{ and } |S(\mathrm{x}_{i+1}^{r}) - S(\mathrm{x}_{i}^{r})| > \delta * S(\mathrm{x}_{i+1}^{r}) \quad (6)$$

$$\mathrm{x}_{i+1} = \mathrm{x}_{i+1}^{l} \quad (7)$$

$$\text{elseif } |S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i}^{l})| > \delta * S(\mathrm{x}_{i+1}^{l}) \text{ and } |S(\mathrm{x}_{i+1}^{r}) - S(\mathrm{x}_{i}^{r})| < \delta * S(\mathrm{x}_{i+1}^{r}) \quad (8)$$

$$\mathrm{x}_{i+1} = \mathrm{x}_{i+1}^{r} \quad (9)$$

else

$$\mathrm{x}_{i+1} = \mathrm{x}_{i} \quad (10)$$

else

$$\text{if } |S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i}^{l})| < \delta * S(\mathrm{x}_{i+1}^{l}) \text{ and } |S(\mathrm{x}_{i+1}^{r}) - S(\mathrm{x}_{i}^{r})| > \delta * S(\mathrm{x}_{i+1}^{r}) \text{ and } S(\mathrm{x}_{i+1}^{l}) < S(\mathrm{x}_{i+1}^{r}) \quad (11)$$

$$\mathrm{x}_{i+1} = \mathrm{x}_{i+1}^{l} \quad (12)$$

$$\text{elseif } |S(\mathrm{x}_{i+1}^{l}) - S(\mathrm{x}_{i}^{l})| > \delta * S(\mathrm{x}_{i+1}^{l}) \text{ and } |S(\mathrm{x}_{i+1}^{r}) - S(\mathrm{x}_{i}^{r})| < \delta * S(\mathrm{x}_{i+1}^{r}) \text{ and } S(\mathrm{x}_{i+1}^{l}) > S(\mathrm{x}_{i+1}^{r}) \quad (13)$$

$$\mathrm{x}_{i+1} = \mathrm{x}_{i+1}^{r} \quad (14)$$

else

$$\mathrm{x}_{i+1} = \mathrm{x}_{i} \quad (15)$$

3.3.2C Hand contour detection

For gestures that use the hand contour, it is required that the hand contour is obtained in a computationally efficient manner. Many contour detection techniques have been discussed in the literature involving the following four major approaches: prior knowledge [71], morphology [72], level set [73], and active contour model [74].

In [67], a sequential Monte-Carlo technique based on corner detection and deterministic optimization was proposed to extract a contour. The extraction of the contour depends on the balanced prior constraints on continuity and smoothness. In [68] a morphology based contour extraction algorithm was discussed based on a series of simple morphology transform operations such as dilation, erosion, XOR and union. In [69], a contour algorithm based on the maximum likelihood method was considered. This technique is free of parameter estimation and the segmentation is adaptable to complex connected objects. However, the complexity of the contour extraction is too high for real-time deployment. In [70], the method of active contour model (SNAKE) was shown to be effective. This method does not try to solve the contour problem at one shot. The framework is computationally complex due to its semi-automatic, live-wire edge detection scheme, where the boundary detection problem is formulated as an optimization problem searching for an optimal path between a start-pixel and an end-pixel. The optimal path is the one that generates the minimum cost in traversing from the start-pixel to the end-pixel with the cost being the cumulative cost accumulated going from one pixel to its neighbor. The cost function is a weighted sum of three costs consisting of Laplacian, gradient magnitude, and gradient direction. The computational complexity of this technique does not allow its real-time deployment.

In the devised recognition system, the morphology technique is adopted due to its computational efficiency.

3.3.3 Hand gesture recognition

Two types of hand gestures are considered here. The first type of gestures is motional gestures consisting of seven directional hand gestures of rotation, forward, backward, left, right, up, and down. The second type of hand gestures is finger spelling consisting of six numbers of zero, one, two, three, four, and five. The first type of hand gesture is recognized via the dynamic time warping technique while the second type of hand gesture is recognized via the convex hull technique.

3.3.3A Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is a widely used technique for comparing sequences or time series by searching for optimal alignment. In this section, a brief review of the DTW algorithm is provided.

Let us consider two sequences or time series: $X = \{x_1, x_2, x_3, \ldots x_N\}$ and $Y=\{y_1, y_2, y_3, \ldots y_M\}$ as shown in Figure 3.4. The two sequences can be arranged on the adjacent sides of a grid. The DTW distance calculation starts from the bottom of the grid. Inside each cell, a distance measure is placed for comparing the corresponding elements in the sequences. The best match or the alignment of the points appear in the path through the grid which minimizes the total distance between them. The process of computing the overall distance involves calculating all possible routes through the grid and for each one computing the overall distance. The overall distance of DTW is the minimum of the sum of the distances between the individual elements on a path divided by the sum of the weighting function which is used to normalize the path length.

A warping path of the DTW algorithm satisfies the following three conditions:

- Boundary: The first and last elements of X and Y are matched to each other. The warping path starts and finishes in the diagonally opposite corner of the accumulated cost matrix and will possibly turns back on itself.
- Continuity or step size: This condition restricts the allowable steps in the warping path to adjacent cells. The path advances one step at a time. Both i and j index only increase by at most 1 on each step along the path.
- Monotonicity: The path will not turn back on itself. Both i and j index either stay the same or increase. Figure 3.4 shows an example of the boundary condition.

The optimal warping path is calculated by satisfying the constraints given above with minimal cost. The foregoing constraints allow one to restrict the moves which can be made from any point in the path and so limit the number of paths to be considered. Instead of calculating all the possible routes in the grid which satisfy the above conditions, the DTW algorithm keeps track of the cost of the best path to each point in the grid, and this makes the algorithm effective.

3.3.3B Motional gesture recognition

As was reported in [71], although disparity can provide the hand depth information using a stereo camera, it loses its sensitivity when the hand is hold far from the camera. Here, for forward and backward movements, the contour area variance is used due to its simplicity. For the other motional hand gestures, the Dynamic Time Warping (DTW) algorithm is used as this algorithm is capable of generating the dynamic distance between an unknown gesture signal and a set of reference gesture signals in a computationally efficient manner while coping with different speeds of motional hand gestures. The sample gesture signal comes from the seed point

of the CamShift tracking. More details of the DTW algorithm are discussed in [72] [73]. The warping distance in the DTW algorithm utilized is the sum of Euclidean distances of the time series for three dimensions.

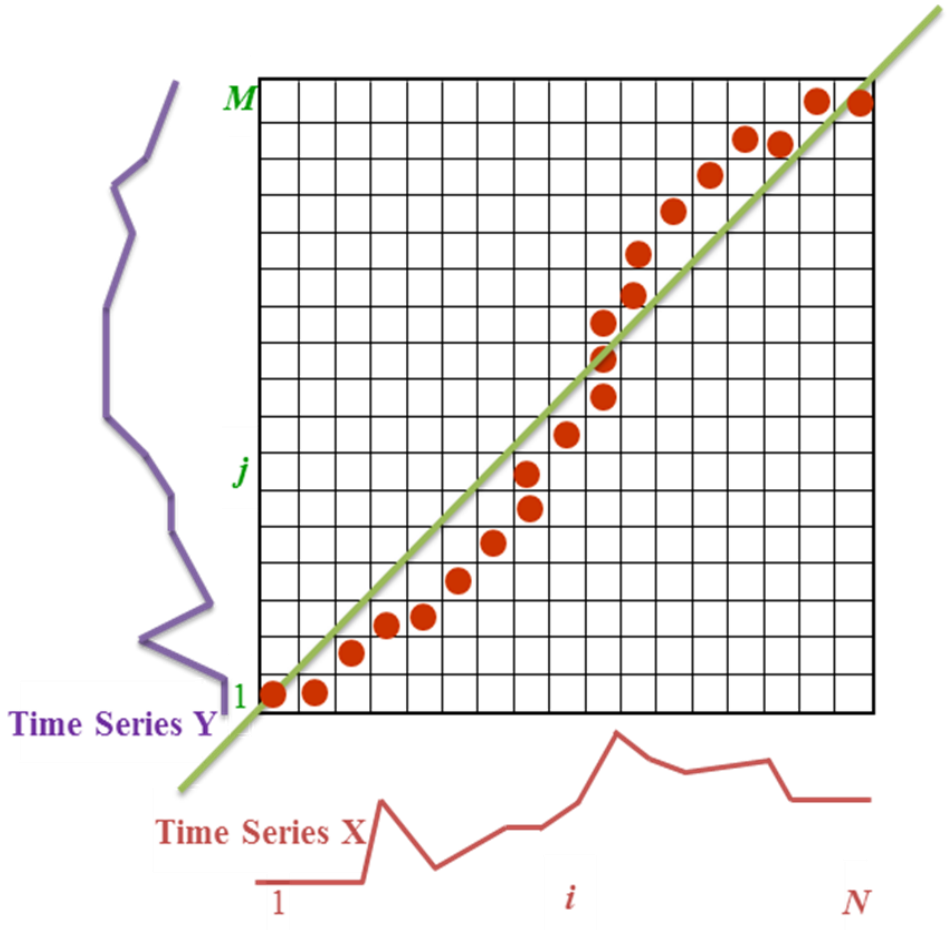

Figure 3.4. DTW grid example

A tree structure is used to indicate the priority level of the hand gestures (see Fig 3.5). For rotation hand gesture, the circumscribed angle of the seed point in the CamShift tracking is used to serve as the motional signal of "rotation" gesture at the highest priority level followed by "forward" and "backward" gestures. The disparity of the seed point is used as the motional signal for these gestures. An actual sample "rotation" signal and its corresponding reference signal are

shown in Figure 3.6. The gradient or position difference of the seed point between consecutive frames is considered to be the motional signal of “left” gesture. An actual sample “left” signal and its corresponding reference signal are shown in Figure 3.7. The gesture for testing DTW recognition appears smooth and without any halting. The range of gesture speeds for DTW is 2-4s.

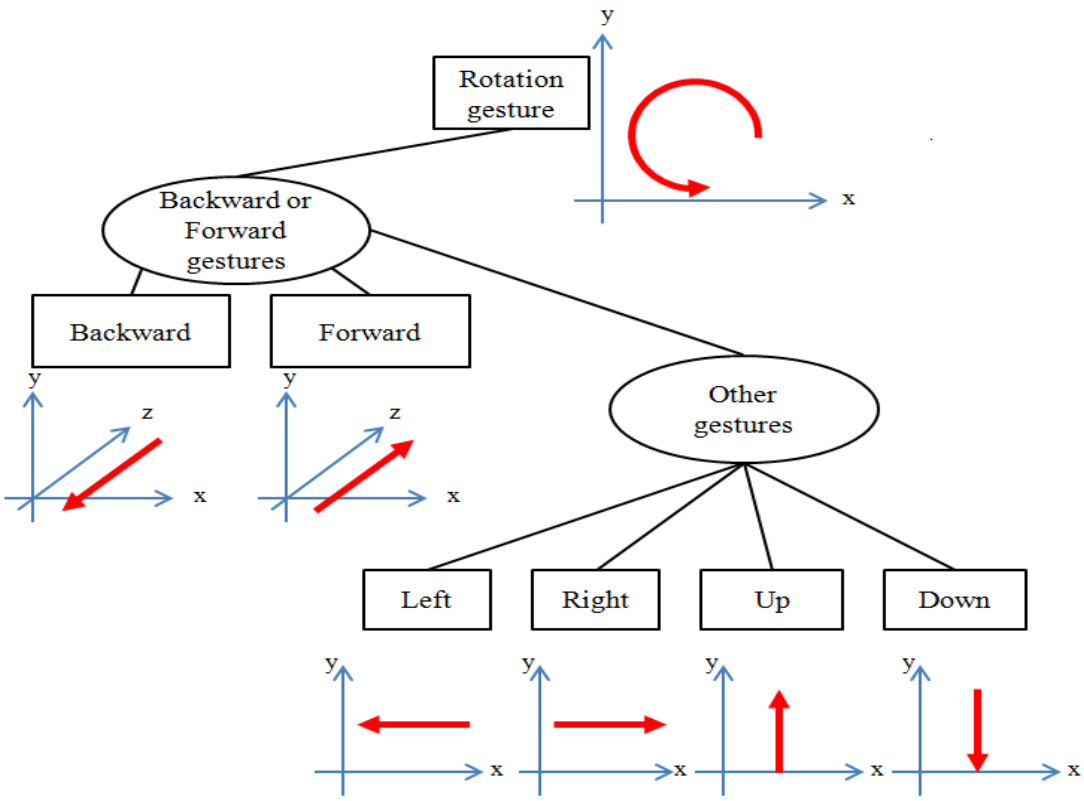

Figure 3.5. Priority level of the hand gestures

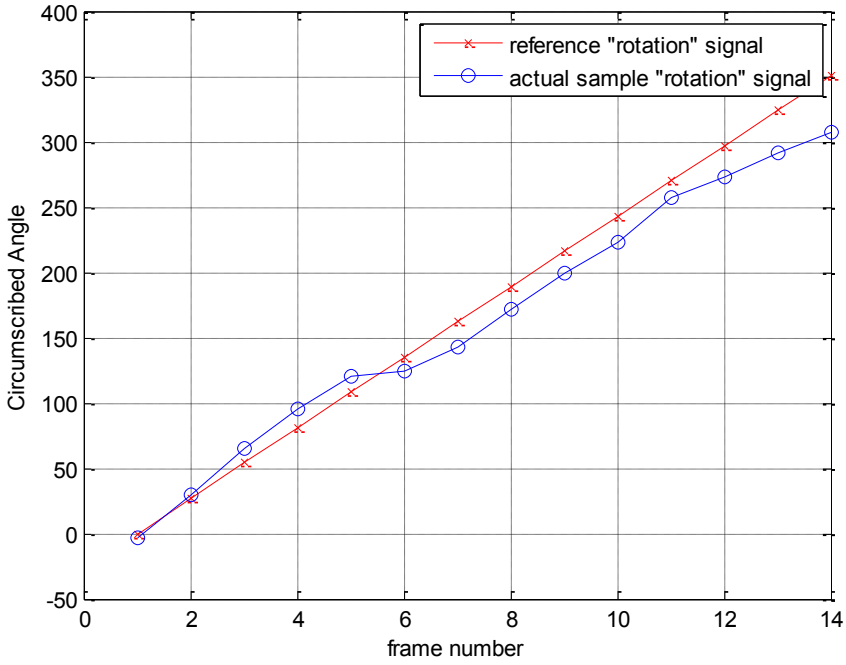

Figure 3.6. A sample “rotation” signal and corresponding reference signal

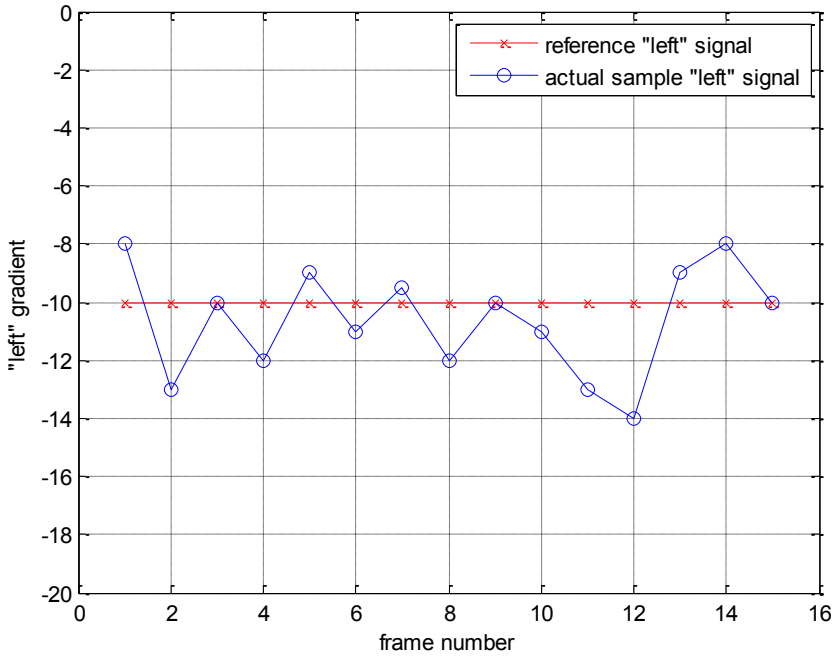

Figure 3.7. A sample “left” signal and corresponding reference signal

3.3.3C Finger spelling recognition

Finger spell recognition is done by first going through the hand contour extraction. Then, the convex hull of the detected contour is used to determine the number of finger tips as reported in [74]. From number 2 to number 5, the recognition can be achieved by the number of the defects from the convex hull. For instance, if there are n complete convex hulls in a hand contour, it implies that there are n+1 fingers. However, this rule does not hold for numbers 0 and 1. The area of the convex hull is thus used instead. As per Equation (16), the area of the convex hull S_{convex} is compared with the contour area $S_{contour}$,

$$\text{Number} = \begin{cases} 0 & S_{convex} \le \alpha * S_{contour} \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

where α denotes a parameter related to the camera distance range where gestures are made. The finger spell recognition requires a relatively high resolution of the hand contour. If the distance range is considered to be too far from the camera, the recognition will suffer due to images having low resolution. If the distance range is considered too close to the camera, the recognition will also suffer, this time due to too much variation in the hand contour. The experimentations done indicated that the following α's provided relatively consistent outcome: α = 10% for 15-35cm camera distance range and 20% for 10-15cm camera distance range. Sample finger spell contours and recognized numbers are shown in Figure 3.8.

Also, based on extensive experimentations, the following operating distance ranges and corresponding α were found to match well: α = 10% for 15-35cm distance from the camera and 20% for 10-15cm distance from the camera.

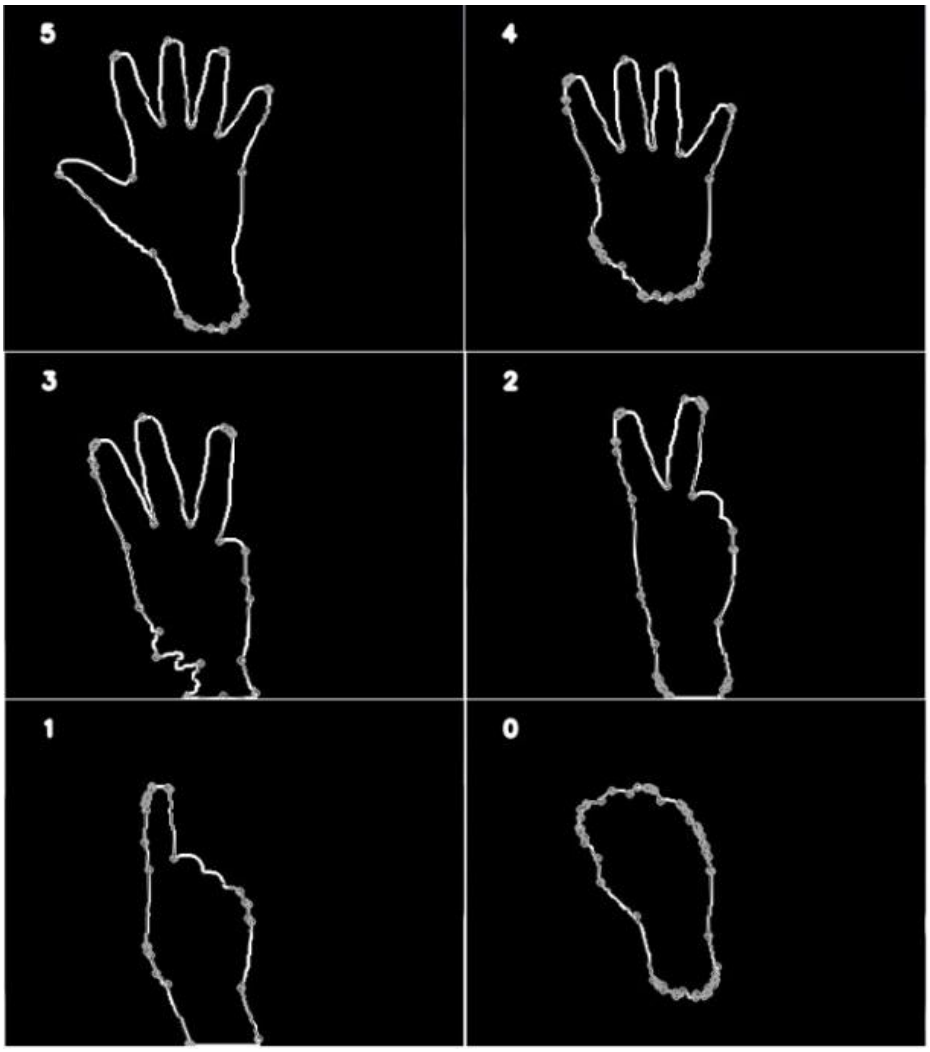

Figure 3.8. Finger spelling contours and recognized numbers

3.4 PREVIOUS WORKS ON DUAL-MODALITY SENSORS FUSION

In previous works on sensor fusion, a number of methods and algorithms have been used including Central Limit Theorem, Kalman filtering, Bayesian network, Dempster-Shafer and Hidden Markov Model. In [75], a probability model using Central Limit Theorem was proposed to detect the location of targets. The independent Poisson random variables were summed over from different infrared cameras. In [76], a real-time navigation system consisting of electro-

optical stereo sensors and Inertial Measurement Unit (IMU) sensors was used to render views from a 3D graphic model. In [77], a road matching strategy based on Dynamic Bayesian Network was utilized for vehicle localization and road matching. This algorithm merged the information of a GPS device and a GIS device to improve the performance of the localization estimation. In [78], a multi-modality biometric identification system based on face and ear bio-information was presented by using Garbor wavelet features and the Dempster-Shafer theory to achieve a robust recognition. In [79], the HMM was used to recognize human actions by fusing data from multiple cameras.

3.5 DUAL-MODALITY SENSORS FUSION

Inertial body sensor and Kinect each has its own shortcomings. Inertial body sensor suffers from long-term drift and Kinect sensor cannot cope with occlusion and its data reliability is lost in the presence of fast movements. In the dual-modality sensors approach in this work, a framework is introduced in order to fuse the data from a wireless inertial body sensor and a vision-based Kinect sensor. The motivation behind this data fusion is to utilize the strengths of these sensors at the same time in order to achieve robustness for hand gesture recognition. By using both of the sensors, a total of 9 signals are generated: 3 Kinect depth coordinate signals of the hand skeleton location and 6 acceleration and gyro signals from the wireless inertial body sensor worn on a subject's wrist as illustrated in Figure 3.10. An example position signal from the Kinect camera and an example acceleration signal from the inertial sensor are shown in Figure 3.9. All of the signals are then fed simultaneously into a multi-HMM classifier to recognize hand gestures.

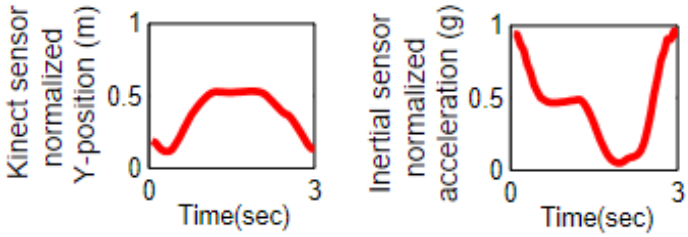

Figure 3.9. Example signals from Kinect depth camera (left) and wireless inertial body sensor (right)

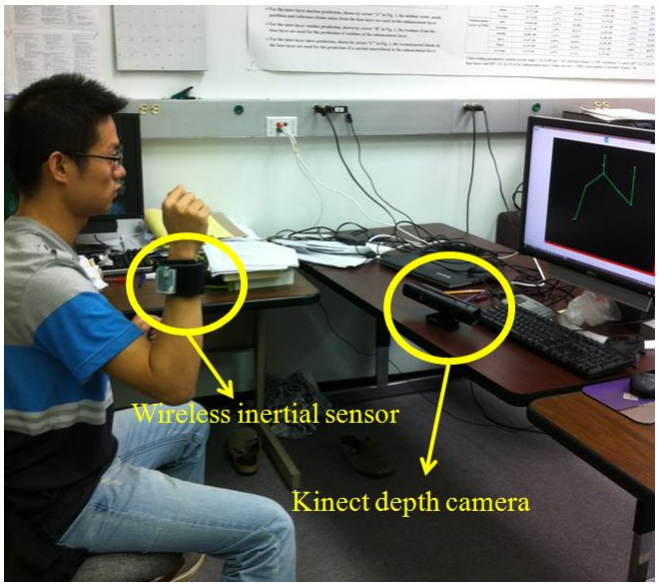

Figure 3.10. Two differing modality sensing for hand gesture recognition

3.5.1 Resampling and filtering

The sampling rates of the Kinect and inertial sensor used are 30 Hz and 200 Hz, respectively. Thus, in order to fuse the data from these two sensors, the inertial sensor data is down-sampled to match the sampling frequency of the Kinect. Because of the presence of various noise sources in an actual operating environment, jitters often appear in the Kinect skeleton signal as well as in the inertial signal. A moving average window is thus used in order to reduce jitters in the signals. After carrying out extensive experimentations, it was found that a

moving window of size between 9 and 19 generates a substantial reduction of jitters in the signals. Figure 3.11 shows an example of the raw and filtered signals from the Kinect and inertial (IMU) sensors.

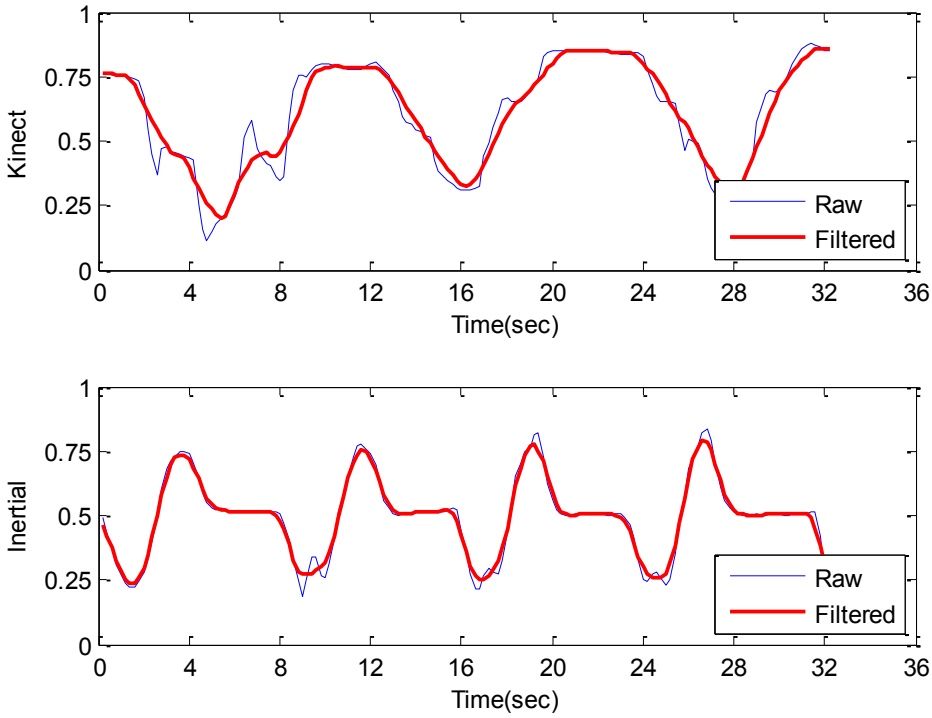

Figure 3.11. Raw signal vs. filtered signal, top: Kinect, bottom: inertial sensor

3.5.2*HMM classifier*

Markov chain is a widely used model to cope with a random sequence of a finite number of states. In this work, arrays of consecutive gesture coordinates in Kinect and arrays of angular velocities of an inertial sensor are considered to form a Markov chain where the true states of the model $S = \{S_1, S_2, \ldots, S_M\}$ are hidden in the random signal sequences which are not directly observable. This type of Markov chain model is known as Hidden Markov Model (HMM).

HMMs aim at building a probability model to represent the true states S from an actual observation sequence $O = \{O_1, O_2, \ldots, O_T\}$ and a state sequence $Q = \{q_1, q_2, \ldots, q_T\}$ together with the probability of the observation sequence O.

The HMM model characterizes a state transfer probability distribution A and an observation symbols probability distribution B. Given an initial state matrix π, an HMM is described by the triplet $\lambda = \{\pi, A, B\}$. Since gesture recognition involves temporal signal sequences, a left-right HMM topology is adopted here, see Figure 3.12.

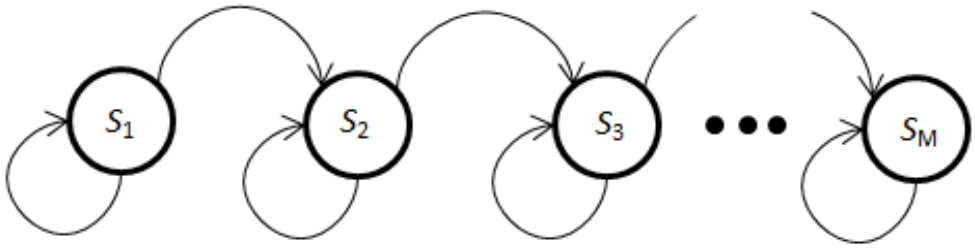

Figure 3.12. Left-right HMM topology

A brief mention of the HMM equations are provided in this section. More details are available in [16]. Suppose a random sequence $O = \{O_1, O_2, \ldots, O_T\}$ is observed; let $V = \{v_1, v_2, \ldots, v_T\}$ denote all possible outcomes and let $S = \{S_1, S_2, \ldots, S_M\}$ denote all HMM states with q_t representing the state at time t, where T indicates the number of time samples. Then, the HMM probability matrices are as follows:

$$\pi = \{p_i = P(Q_1 = S_i)\},\ 1 \le i \le M; \tag{17}$$

$$A = \{a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)\},\ 1 \le i, j \le M; \tag{18}$$

$$B = \{b_j(k) = P(O_t = v_k \mid q_t = S_j)\},\ 1 \le j \le M,\ 1 \le k \le T; \tag{19}$$

$$\text{where} \sum_{i=1}^{M} \pi_i = 1,\ \sum_{j=1}^{M} a_{ij} = 1 \text{ and } \sum_{k=1}^{T} b_j(k) = 1 \tag{20}$$

For the training of HMM, first its parameters need to be initialized. Among all the initialization matrices, the most important initialization is the transition matrix A where the constraints to control the initial transitions are set. By zeroing out all the nonadjacent probabilities in this matrix, the state transitions are made limited to the sequence of adjacent states representing a hand gesture. That is to say, for the application under consideration here, the prior is set to constrain all possible state transitions to only occur from left-to-right and between two adjacent states. A typical initial transition matrix A is thus formed to be

$$A = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

Based on the initialization matrices, let $O = \{O_1, O_2, \ldots, O_T\}$ be the observation sequence of a hand gesture, $Q = \{q_1, q_2, \ldots, q_T\}$ be the corresponding state sequence with the probability of the observation sequence O given by $P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda)$. According to the Baum-Welch algorithm [16], the probability $P(O|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} a_{q_3 q_4} \cdots a_{q_{T-1} q_T}$ can get calculated towards updating λ. Since $P(O, Q|\lambda) = P(O|Q, \lambda)P(Q, \lambda)$, one gets

$$P(O|\lambda) = \sum_Q P(O|Q,\lambda)\mathrm{P}(Q,\lambda) = \sum_{q_1,q_2,\ldots,q_T} \pi_{q_1}\, b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_T}(O_T) \tag{22}$$

To update the current model $\lambda = \{\pi, A, B\}$, let the updated model be $\bar{\lambda} = \{\overline{\pi}, \overline{A}, \overline{B}\}$. For estimating the model $\bar{\lambda} = \{\overline{\pi}, \overline{A}, \overline{B}\}$, let the probability of the joint event that $O_1, O_2, \ldots, O_t$ is observed be $\alpha_t(i)$, thus $\alpha_t(i) = P(O_1, O_2, \ldots, O_T, q_T = S_i|\lambda)$. In a backward way, let $\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T, q_T = S_i|\lambda)$. The probability being in state S_i at time t and state S_j and time t+1 is thus given by

$$\xi_t(i,j) = P\left(q_t = S_i, q_{t+1} = S_j \middle| O, \lambda\right) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} \tag{23}$$

Given $\gamma_t(i)$ as the probability of state S_i at time t, one gets $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$, $\bar{\lambda} = \{\overline{\pi}, \overline{A}, \overline{B}\}$, where

$$\overline{\pi}_i = \gamma_t(i) \tag{24}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{25}$$

$$\overline{b_j}(k) = \frac{\sum_{t=1, O_t = v_k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{26}$$

Considering a very small threshold value, e.g. $\varepsilon = 10^{-6}$, or if $log\{ P(O|\lambda)\} - log\{P(O|\bar{\lambda})\} < \varepsilon$, the training can get terminated. The flow chart of the training process is shown in Figure 3.13.

For testing, a test sequence is fed into several trained HMM models each corresponding to a hand gesture in order to calculate the likelihood probabilities. Then, a high (e.g., 95%) confidence interval is applied to the probabilities to classify the sequence. Let μ and σ represent the mean and variance of the likelihood probabilities. For the 95% confidence interval, whenever none of the probabilities is larger than $\mu + 1.96\frac{\sigma}{\sqrt{n}}$ (where $n = 5$ when testing based on Microsoft MSR hand gesture dataset, n = 10 when testing on $1 Recognizer hand gesture dataset), the sequence is rejected and the gesture is considered to be a NOT-DONE-RIGHT gesture. If the

sequence is not rejected, the gesture with the maximum probability is considered to be the recognized gesture, see Figure 3.14.

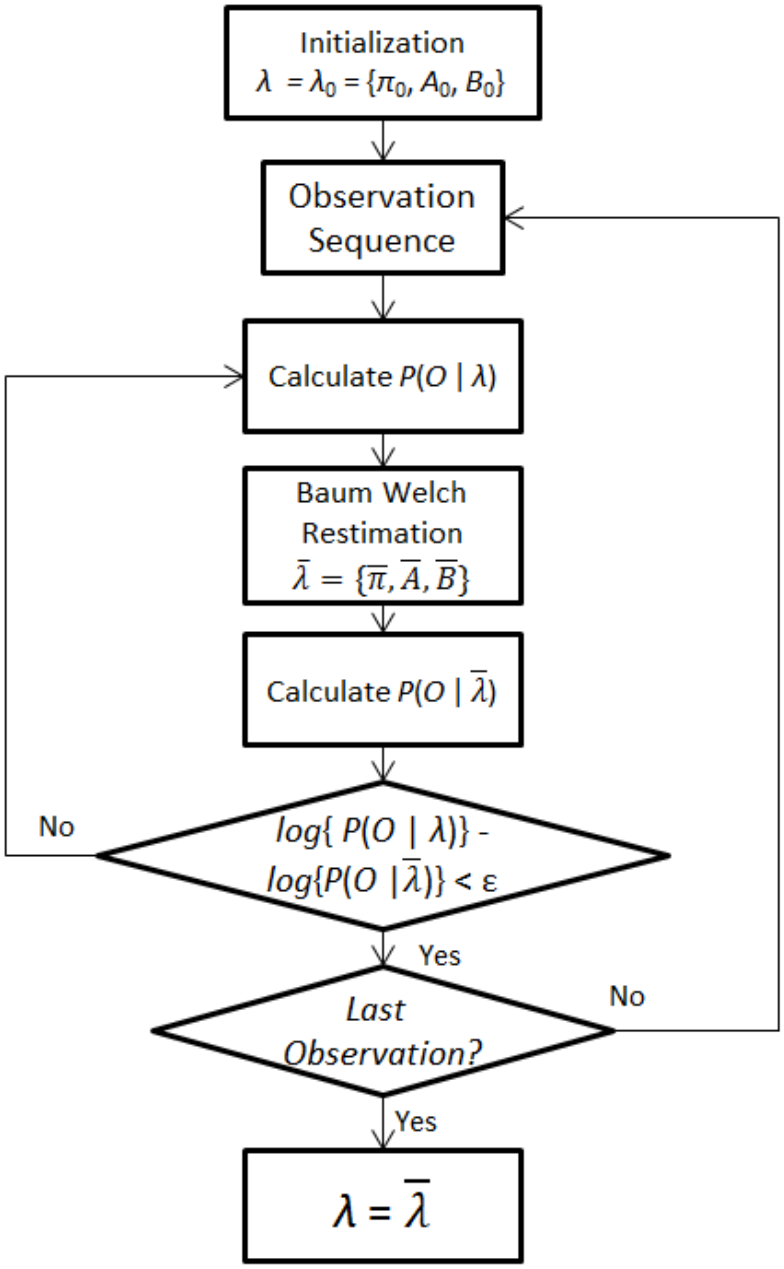

Figure 3.13. Flowchart of HMM training

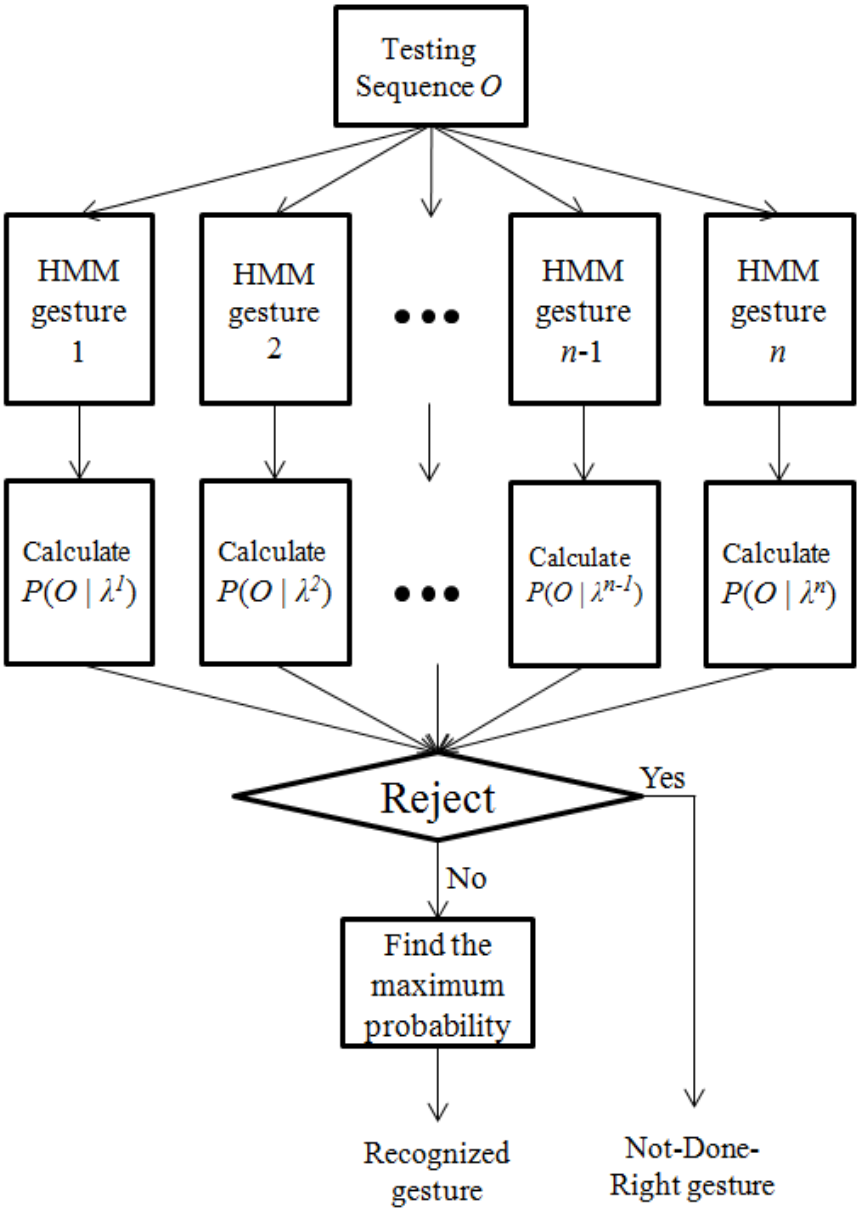

Figure 3.14. Flowchart of HMM testing or recognition

3.6 MULTI-HMM CLASSIFICATION

During operation or testing the HMM models, the variance of the likelihood probabilities is sometimes not adequate enough to recognize the hand gestures. The small variance can be

traced back to the intrinsic drift error of the inertial sensor or to the unreliable Kinect data due to rapid hand movements. In these cases, HMMs could fail and the likelihood probabilities may not be distinct enough and thus false negative would occur. To enhance the differences among the likelihood probabilities and to decrease false negatives, an improvement of the previous solution is done in this section by utilizing a multi-HMM classification [90] for recognition of single hand gestures.

The improvement is made by combining the decisions of multiple HMM classifiers to form a mixture model. Often, the data captured by one type of sensor does not capture all the variations of a hand gesture. However, by using differing modality sensors within a multi-HMM classification framework, it is hypothesized that a more robust recognition can be achieved under realistic conditions. Multi-HMM classification is not a new concept and has been previously applied to many applications, e.g. text recognition [80], handwriting recognition [81], finger-print recognition [82], speaker recognition [83]. However, this work is the first time such a classification approach is applied to hand gesture recognition based on two differing modality sensors. More specifically, the fusion approach introduced in this dissertation differs from all the previous approaches not only by using a multi-HMM classification but also by using the data from both a depth camera (Kinect) and a wearable inertial sensor. Another important aspect of this paradigm is that the computational complexity of the recognition pipeline is kept low leading to its real-time implementation.

In what follows, it is discussed how the multi-HMM classification increases the robustness of recognition. An HMM model characterizes a state transfer probability matrix A and an observation symbols probability matrix B. Given an initial state matrix π, an HMM is

described by the triplet $\lambda = \{\pi, A, B\}$. Here, a left-to-right HMM topology is adopted since hand gesture recognition involves temporal signal sequences. Let $O = \{O_1, O_2, \ldots, O_T\}$ be the observation sequence of a hand gesture, where T denotes the number of time samples. The theory of HMM is well established and the details on HMM can be found in many references, e.g. [16].

In spite of intrinsic drift errors associated with inertial sensors and Kinect errors due to rapid hand movements, the improvement made to increase the overall correct recognition rate is reported next.

Similar characteristic input signals of coordinates, acceleration, and angular gyro are clustered and fed into three component HMM classifiers, each classifier generating its own likelihood probability as shown in Figure 3.15. All likelihood probabilities from the component classifiers are then multiplied by equal weights and are pooled together to generate an overall probability $P(O|\lambda)$ for the input signals.

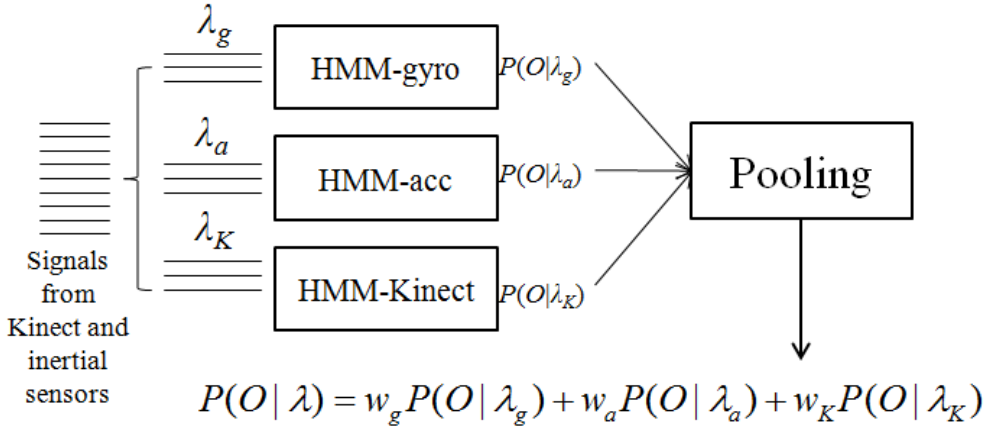

Figure 3.15. Framework of the multiple HMM classification

Each hand gesture data consists of 9-dimensional signals (3 dimensions for angular gyros, 3 dimensions for accelerations, and 3 dimensions for Kinect hand skeleton coordinates). The models considered are denoted by $\lambda_{g0} = \{\pi_{g0}, A_{g0}, B_{g0}\}$, $\lambda_{a0} = \{\pi_{a0}, A_{a0}, B_{a0}\}$ and $\lambda_{K0} = \{\pi_{K0},$

A_{K0}, $B_{K0}\}$ representing gyro, accelerometer and Kinect HMM models, respectively. The parameters of these models are then estimated according to the Baum-Welch algorithm [11]. During operation or testing, $P(O|\lambda)$ is represented by the three likelihood probabilities $P(O|\lambda_g)$, $P(O|\lambda_a)$ and $P(O|\lambda_K)$. The gesture with the maximum average of the three likelihood probabilities is then considered to be the recognized gesture. As a result of using a multi-HMM in this manner, the difference of the probability likelihoods gets diminished or the discriminatory power gets increased.

3.7 SUMMARY

Three new real-time hand gesture recognition systems have been discussed in this chapter using different fusion approaches. First, based on a pair of stereo images, a rule-based framework merging the information from the left and right images of a stereo image pair was introduced. Second, a data fusion approach to hand gesture recognition based on the probabilistic HMM classification involving the two sensors of inertial body sensor and Kinect depth sensor was introduced. Third, a multi-HMM classification was considered for the two differing modality sensors in order to improve the recognition performance of this real-time hand gesture recognition system.

CHAPTER 4

RECOGNITION RESULTS AND DISCUSSION

In this chapter, the results of various experiments conducted are presented and discussed to see how well the developed hand gesture recognition systems perform. First, the DTW hand gesture recognition rate based on different distances of time series are compared. Then, for the stereo fusion approach, the results of hand detection in different lighting conditions, as well as motional hand gesture recognition confusion matrices, finger spelling recognition confusion matrices are presented when using a single image and when using a pair of images. In the third part, a cross-modality comparison between the two dual-sensors approaches is presented. Finally, the classification outcome of using one HMM model and a multi-HMM model are compared.

4.1 DTW RECOGNITON RATE BASED ON DIFFERENT DISTANCES

DTW allows comparing two temporal sequences, a sample sequence X: (x_1, x_2, ..., x_n) and a testing sequence Y: (y_1, y_2, ..., y_m), based on a local distance measure. In this section, the recognition rate based on different distance measures consisting of Manhattan distance (L_1), Euclidean distance (L_2) and p-norm distances (L_p) are compared.

Manhattan distance, see Equation (27), also called Taxicab distance indicates the distance a taxi has to drive in a rectangular street in Manhattan to get from an origin to a

destination. The L_1 norm of the vector $\boldsymbol{d}$ is the sum of the absolute values of all the terms, p is the number of warping path steps, normally it is bigger than n and m,

$$\|\boldsymbol{d}\| = \sum_{i=1}^{p} |x_i - y_i| \tag{27}$$

For the Euclidean distance calculation, the local distance between a sample sequence and a testing sequence is computed as per Equation (28):

$$\|\boldsymbol{d}\| = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \tag{28}$$

The p-norm distances, where q>2, is computed as per Equation (29):

$$\|\boldsymbol{d}\| = \left(\sum_{i=1}^{p} |x_i - y_i|^q\right)^{\frac{1}{q}} \tag{29}$$

Note that for q=1, the Manhattan distance is resulted and for q=2,the Euclidean distance is resulted. Considering the tradeoff between computational complexity and recognition rate, Euclidean distance was found to be the best choice here. As can be seen from Equations (27)-(29) and Table 4.1, the computational complexity of the distance measures increases proportionally with q, while the hand gesture recognition rate does not increase proportionally with q.

Table 4.1. Computational complexity of DTW using different distances

	Manhattan distance	Euclidean distance	p-norm distances
Complexity	$O(p)$	$O(p^2)$	$O(p^q)$

As can be seen from Figure 4.1, the recognition rate achieved by using the Euclidean distance is competitive with the other distances while having a relatively lower computational

complexity. Moreover, since the mapping between the components of two temporal sequences is fixed, the Manhattan distance is very sensitive to noise and time misalignment and thus are not suitable to recognize relatively complex gesture such as DrawX.

One gesture set consisting of the five single hand gestures in the Microsoft Action Dataset [14] was used to examine the recognition outcome based on the DTW algorithm. There are 5 single hand gestures in the Microsoft Action Dataset and the gestures are illustrated in Figure 4.2. Ten subjects were asked to perform these five gestures 30 times in front of different backgrounds. Different backgrounds included different scenes appearing in different lighting conditions including outdoor day light, indoor florescent and indoor incandescent lights. Each subject performed the gestures at different speeds which were timed to last between 1 to 3 seconds. The variance of Figure 4.1 is based on ten subjects. q=5 for the p-norm distance was used.

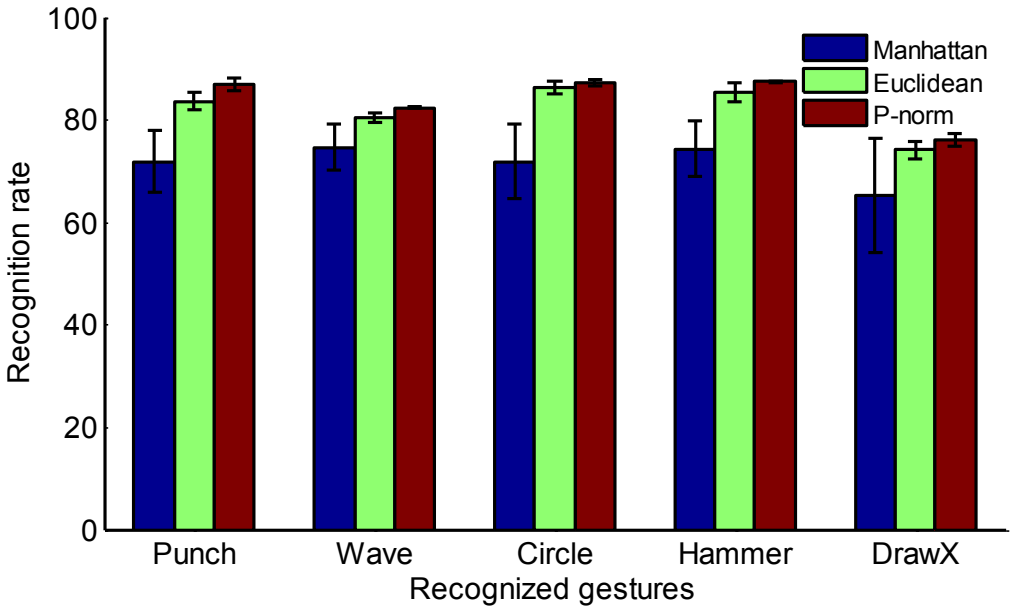

Figure 4.1. Hand gesture recognition rate (%) and variance based on different distances.

Figure 4.2. Single hand gestures in the Microsoft Action Dataset: “wave”, “hammer”, “punch”, “drawX”, “circle”

4.2 COMPARISON OF SINGLE AND STEREO CAMERA

This section covers the experiments carried out to show the increase in robustness when the information from two cameras were used versus using the information from a single camera. The algorithm is written in C which was run on a PC with a dual core 2.1GHz processor. The input images were captured with the stereo webcam Novo Minoru, which is an inexpensive stereo webcam generating low resolution images of size 640*480 shown on the left side of Figure 4.3. Additional stereo images were examined using a Fuji stereo digital camera shown on the right side of Figure 4.3.

Figure 4.3. Novo Minoru and Fuji stereo digital camera

Table 4.1 provides a comparison of the hand detection outcome using a single image versus a pair of stereo images under different lighting conditions. As can be seen from this table, both the single and stereo image approaches achieved a frame rate of 31-32 frames per second (fps) or about 30ms per frame. However, as shown in Table 4.2, by merging the information from the left and right images, the average percentage detection rate was considerably improved (by nearly 60%). It is worth noting that in the presence of adequate lighting (florescent), both single and stereo images achieved a relatively higher detection rate. While in the presence of inadequate or outdoor lighting (low light and sunlight), the utilization of stereo images outperformed single images due to the rule-based fusion of stereo images.

In the experiments carried out, 50 cases of each movement ("Left", "Right", "Up", "Down", "Forward", "Backward" and "Rotation") were considered under various lighting and background conditions. Tables 4.3 and 4.4 provide the recognition confusion matrices when using single images versus when using stereo images. As can be seen from Table 4.4, by combining the information from the left and right cameras, in particular for backward and forward motions, the overall recognition rate was significantly improved. Also, as noted in Figure 4.4, the variance when using single images is larger than when using stereo images. This is caused by merging the gesture information based on stereo images, the robustness is enhanced,

in particular for the gestures "Forward" and "Backward". The use of a single camera does not allow one to detect the depth information. In this comparison, the growing and shrinking of the hand blob was used to represent the "Forward" and "Backward" gesture cases.

For motional hand gestures, the average recognition rate reached 93% when using stereo images as compared to 66% when using single images. Notice that there was a 4% of no detection. This was attributed to the low resolution of the captured hand images.

The finger spell recognition comparison when using single images versus when using stereo images is provided in Tables 4.5 and 4.6. For finger spell recognition, the average recognition rate reached 92% when using stereo images as compared to 62% when using single images. For the finger spelling experiments, 100 cases of each finger spelling number were considered under various lighting conditions. As can been seen from Tables 4.5 and 4.6, by merging the hand contour of the stereo images, the recognition rate was improved. Since the shape (convex hull) and the orientation of the hand contour highly depend on the camera viewing angle and the lighting conditions, the features of finger spelling are not always reliable for recognition in the case of a single camera. The application of stereo cameras can alleviate some of the problems. Since the viewing angles of a pair of stereo cameras are different, the shape and extracted hand contours from the two cameras act in a complementary manner. Also, as seen from Figure 4.5, the variance when using stereo images was smaller than when using single images.

Table 4.2. Comparison of hand detection rates when using single images versus pairs of stereo images

LIGHTING CONDITION	Single image		Stereo images	
	Detected frames	Frame rate	Detected frames	Frame rate
FUJI_FLORESCENT	105/245	33	230/245	31
MINORU_FLORESCENT	137/350	33	329/350	31
MINORU_FLORESCENT_LOWLIGHT	82/315	31	215/250	30
MINORU_SUNLIGHT	56/200	32	178/200	32
FUJI_INCANDESCENT	144/400	34	368/400	32
MINORU_COMPLEXBACK_ FLORESCENT	62/295	31	264/295	30
MINORU_INCAND_LOWLIGHT	52/275	31	232/275	30
MINORU_INCAND	64/265	32	236/265	31
Average detection rates	26%	32	89%	31

In Table 4.7, the total processing time for all the components in the recognition system is listed, which is approximately 30ms per frame. Note that the online color calibration took 1sec but it is not included in the table since it is done only once at the beginning when the system is turned on.

Table 4.3. Motional hand gesture recognition confusion matrix when using single images

Hand gesture recognition rates (%)		Recognized Gesture							
		L	RI	U	D	F	B	RO	No detection
Actual Gesture	Left(L)	76	4	2	4	2	4	2	6
	Right(RI)	6	72	4	2	4	2	2	8
	Up(U)	2	4	76	2	4	0	2	10
	Down(D)	4	2	4	74	2	4	4	6
	Forward(F)	2	6	4	10	42	20	12	4
	Backward(B)	2	4	2	14	18	48	6	6
	Rotation(RO)	2	0	2	2	4	6	74	10

Table 4.4 Motional hand gesture recognition confusion matrix when using stereo images

Hand gesture recognition rates (%)		Recognized Gesture							
		L	RI	U	D	F	B	RO	No detection
Actual Gesture	Left	98	0	0	0	0	0	0	2
	Right	0	96	0	0	0	2	0	2
	Up	0	0	96	0	0	0	0	4
	Down	0	0	0	96	2	0	0	4
	Forward	4	2	2	0	86	0	0	6
	Backward	0	2	2	0	0	92	0	4
	Rotation	0	0	4	2	0	0	88	6

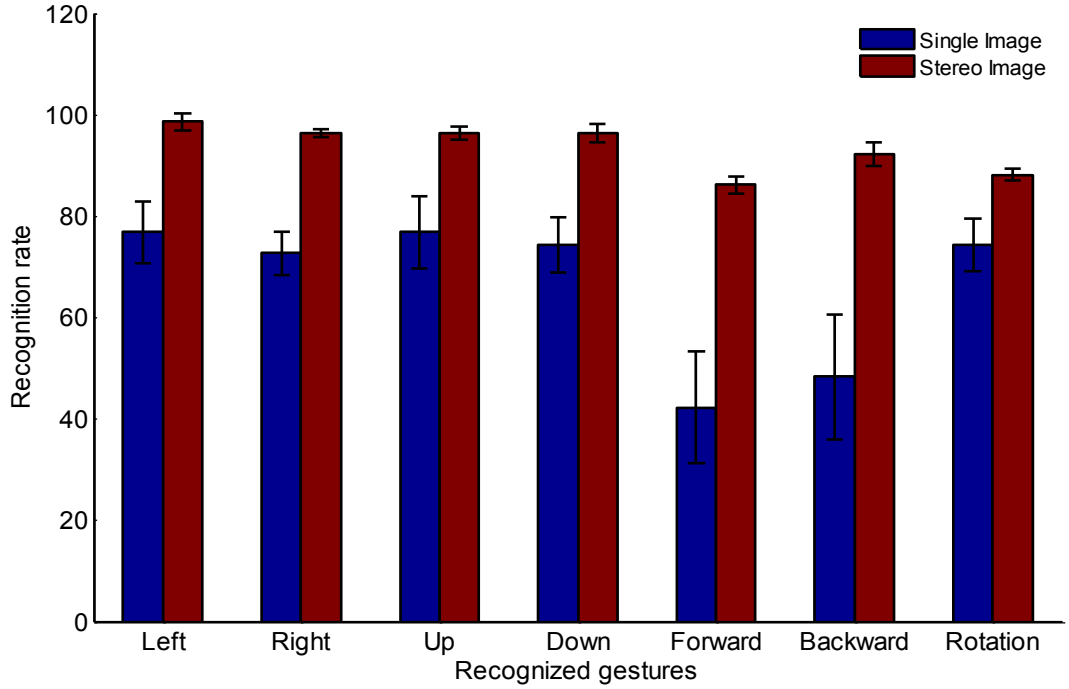

Figure 4.4. Hand gesture recognition rate (%) when using single images versus pairs of stereo images.

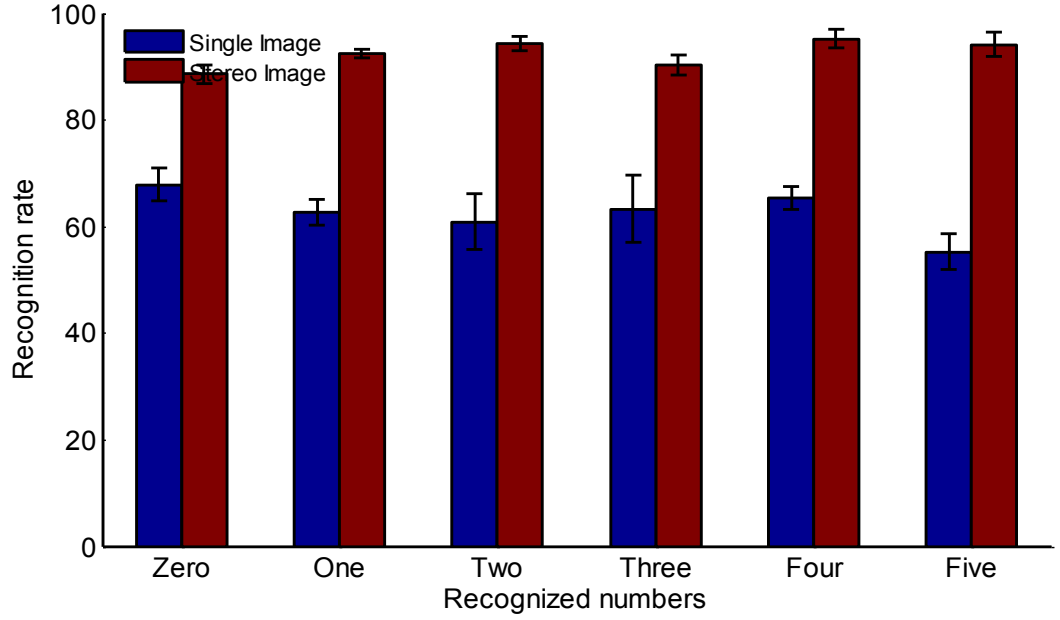

Figure 4.5. Finger spelling recognition rate (%) when using single images versus pairs of stereo images.

Table 4.5. Finger spelling recognition confusion matrix when using single images

Finger spelling recognition rates (%)		Recognized Number					
		Zero	One	Two	Three	Four	Five
Actual Number	Zero	67	15	18	0	0	0
	One	31	62	5	1	1	0
	Two	5	6	60	25	4	0
	Three	6	10	6	63	5	10
	Four	0	10	10	3	65	12
	Five	0	5	6	10	24	55

Table 4.6. Finger spelling recognition confusion matrix when using stereo images

Finger spelling recognition rates (%)		Recognized Number					
		Zero	One	Two	Three	Four	Five
Actual Number	Zero	88	12	0	0	0	0
	One	0	92	4	4	0	0
	Two	0	4	94	2	0	0
	Three	0	0	4	90	3	3
	Four	0	0	0	2	95	3
	Five	0	0	0	3	3	94

Table 4.7. Average and standard deviation processing times of the component of the introduced approach

Components	Processing time (ms)
Color based hand tracking	5±0.5
Gesture recognition	15±0.9
Finger spell recognition	10±1.0
Total processing time per frame	30±2.4

In a different set of experiments, the developed approach was compared to two existing approaches in the literature that have been shown to provide high recognition rates, namely optical flow and HMM.

Table 4.8 provides a comparison of the developed approach with these approaches. As shown in this table, although the recognition rates between the three approaches were more or less comparable, the dual-sensor approach achieved a higher frame rate leading to a real-time throughput.

4.3 CROSS COMPARISON BETWEEN STEREO CAMERA APPROCH AND DUAL-MODALITY APPROACH

In this section, the two approaches of stereo cameras and dual-modality sensors are compared. Two gesture sets were considered for the comparison of two approaches. One gesture set consisted of the single hand gestures in the Microsoft Action Dataset [14] and the other gesture set consisted of the $1 Gesture Recognizer Dataset [34]. There are 5 single hand gestures in the Microsoft Action Dataset and 15 single hand gestures in the $1 Gesture Recognizer

Dataset. The gestures in the Microsoft Action Dataset are illustrated in Figure 4.2. The hand gestures in the $1 Gesture Recognizer Dataset are used to manage and navigate an Opera Web Browser [35]. These gestures are illustrated in Figure 4.6 with the beginning of a gesture indicated by a solid dot.

Table 4.8. Comparison of average and standard deviation recognition and frame rates between two existing approaches and the introduced approach

LIGHTING CONDITION	Optical Flow		HMM		Introduced approach	
	Recognition rate (%)	Frame rate/s	Recognition rate (%)	Frame rate/s	Recognition rate (%)	Frame rate/s
FLORESCENT	91±3.1	3±0.4	92±2.3	4±1.4	94±3.0	31±1.0
SUNLIGHT	84±5.2	2±2.0	87±5.2	6±4.0	89±5.3	32±3.4
INCANDESCENT	87±4.2	3±1.5	91±3.0	3±1.5	89±3.1	31±1.6
Average and standard deviation of recognition rate and frame rate per second	87±4.2	3±1.3	90±3.5	4±2.3	90±3.8	31±2.0

In the stereo recognition system, the 3D hand position is localized by a stereo camera, and thus the centroids of the hand blobs and the hand bounding boxes are used to represent the hand movement. The DTW algorithm is then used to generate a dynamic distance between a hand gesture model signal and an actual hand movement signal. While in the dual-modality recognition system, the temporal hand gestures are considered to be statistical variations in both positions and state transitions among a set of dynamic models. By feeding the trajectories into

trained models, HMMs are used to generate the statistical variations in both the position and transition of hand movements, as well as to segment the gesture stream automatically.

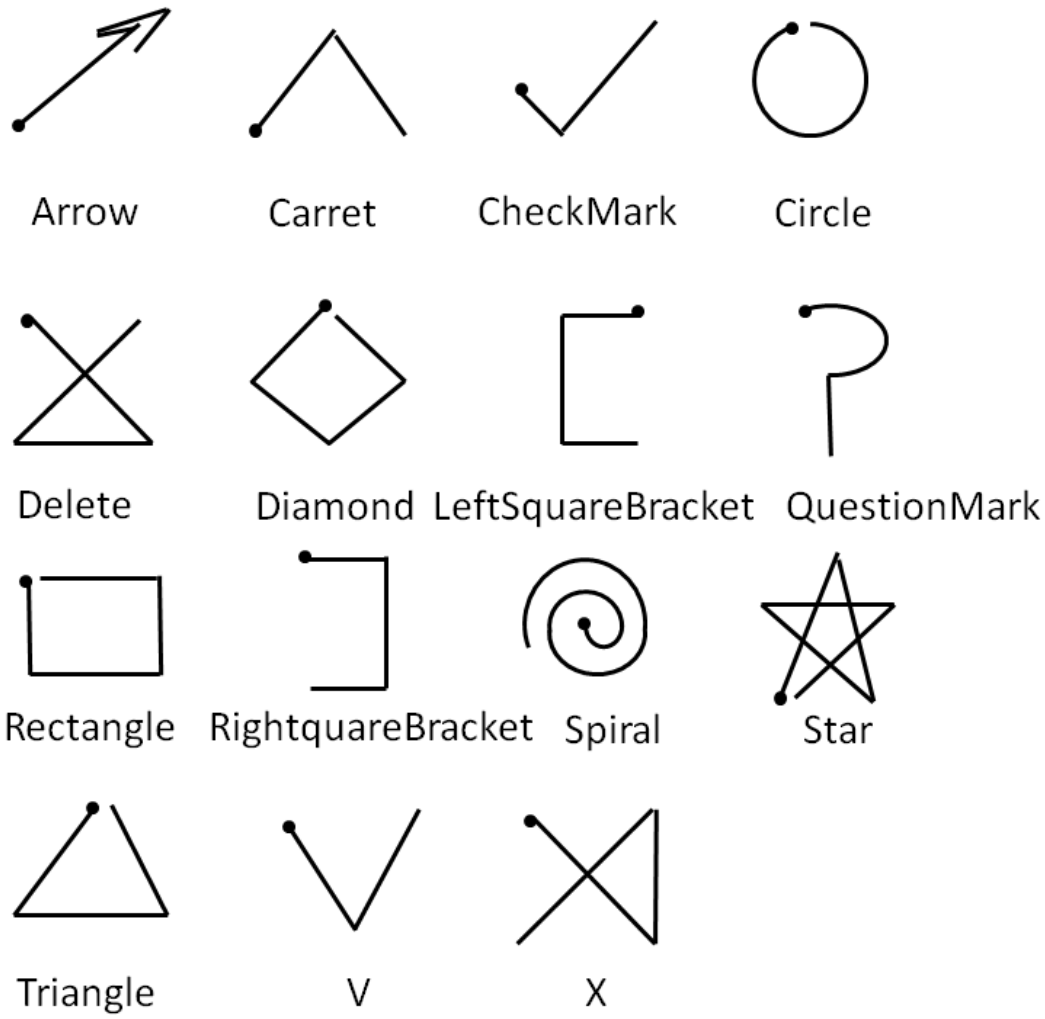

Figure 4.6. Single hand gestures in the $1 Gesture Recognizer Dataset

In the experiments carried out, ten subjects were asked to perform each gesture in the two gesture sets 30 times with different speeds in front of different backgrounds. For the stereo camera fusion system, the DTW distance indicated how well an unknown signal matched a number of template or reference signals. A template or reference signal of a hand gesture was set

up by taking the average of the training sample signals. The signals consisted of the 3-axis {X, Y, Z} gradient signals after merging the information from the stereo images. For the dual-modality sensor fusion, 8-12 HMM states were used. The 3-axis accelerometer and the 3-axis gyroscope signals from the wireless inertial sensor and the 3-axis {X, Y, Z} gradient or position difference signals from the Kinect camera were captured simultaneously in a synchronized manner to form the observation sequence $O = \{O_1, O_2, \ldots, O_T\}$ of the HMM classifiers. The recognition process was repeated 10 times. Each time a different set of 9 training subjects was chosen. The recognition rates obtained were averaged to remove any bias to a particular subject. In addition to performing the hand gestures correctly, incorrect gestures were also performed such that half of them were the same gestures but done in an incomplete way and the other half were random gestures. The incorrectly performed gestures is named "Negative" here while the correctly performed gestures named "Positive".

Table 4.9 shows the performance outcome of the stereo fusion system and Table 4.10 shows the performance outcome of the dual-modality fusion system based on the Microsoft Action gesture set. The experiments included a positive database containing a total of 5*30=150 correctly done gestures and a negative database containing 50 incorrectly done gestures, named "Not-done-right (N)", with 25 gestures done in incomplete ways and 25 done by random hand movements. In these tables, PPV (positive predictive value) and NPV (negative predict value) indicate the recognition rates of correct and incorrect hand gestures, respectively; TP, FP, TN and FN denote true positive, false positive, true negative and false negative, respectively. Table 4.11 and Table 4.12 summarize the performance outcomes of the two fusion systems for the $1 Recognizer gesture set. For this gesture set, the experiments included a positive database

containing 15*30=450 correctly done gestures and a negative database containing 100 incorrectly done gestures.

As an alternative way to show the recognition outcomes of the two fusion systems, Table 4.13 through Table 4.17 provide the confusion matrices obtained, where Table 4.15 provides False recognition rate per subject in the Microsoft Action Dataset when using the dual-modality fusion system. In these tables, the following abbreviations are used for the hand gestures: “Arrow (A)”, “Carret (Ca)”, “Check mark (Ch)”, “Circle (Ci)”, “Delete (De)”, “Diamond (Di)”, “Left square bracket (L)”, “Question mark (Q)”, “Rectangle (Re)”, “Right square bracket (Ri)”, “Spiral (Sp)”, “Star (St)”, “Triangle (T)”, “X”, “V” and “Not-done-right (N)”. Figures 4.7 and 4.8 present the recognition rates when using the stereo fusion system versus the dual-modality sensor system. As can be seen from Figure 4.7, in addition to the recognition rate getting improved by merging the information from the left and right images, the gesture "Punch" and "DrawX" still showed relatively low variance. This is attributed to the gesture "Punch", for which detecting the depth by the stereo cameras is not as accurate as Kinect. For the gesture "DrawX", the dual-modality fusion system tracked speedy hand movements more robustly than those of the stereo cameras. Based on $1 Recognizer Dataset, Figure 4.8 shows the hand gesture recognition rate when using the stereo fusion system versus the dual-modality fusion system.

Table 4.9. Recognition rates of the stereo fusion system for the hand gestures in the Microsoft Action Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	123	27	150	PPV=0.82
Negative	42	8	50	NPV=0.83

*PPV(Positive Predictive Value)=TP/(TP+FP) NPV(Negative Predictive Value)=TN/(TN+FN)

Table 4.10. Recognition rates of the dual-modality fusion system for the hand gestures in the Microsoft Action Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	138	12	150	PPV=0.92
Negative	48	2	50	NPV=0.97

Table 4.11. Recognition rates of the stereo fusion system for the hand gestures in the $1 Recognizer Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	338	112	450	PPV=0.75
Negative	73	27	100	NPV=0.73

Table 4.12. Recognition rates of the dual-modality fusion system for the hand gestures in the $1 Recognizer Dataset

Hand gesture	Recognition outcome		Total	Recognition measures
	True	False		
Positive	407	43	450	PPV=0.90
Negative	91	9	100	NPV=0.91

Table 4.13. Confusion matrix (% recognition rates) for the hand gestures in the Microsoft Action Dataset when using the stereo fusion system

	wave	hammer	punch	drawX	circle	N
wave	80	3	3	5	8	1
hammer	2	85	3	5	2	3
punch	4	6	83	2	3	2
drawX	5	4	3	76	7	5
circle	6	2	1	4	86	1
N	1	1	5	9	1	83

Table 4.14. Confusion matrix (% recognition rates) for the hand gestures in the Microsoft Action Dataset when using the dual-modality fusion system

	wave	hammer	punch	drawX	circle	N
wave	92	1	1	5	1	0
hammer	5	91	2	2	0	0
punch	3	5	91	0	0	1
drawX	0	0	6	88	6	0
circle	1	0	0	0	99	0
N	0	1	1	1	0	97

Table 4.15. False recognition rate (%) per subject in the Microsoft Action Dataset when using the dual-modality fusion system

Subject#	1	2	3	4	5	6	7	8	9	10
False recognition rate	6	8	3	11	7	5	8	8	4	10

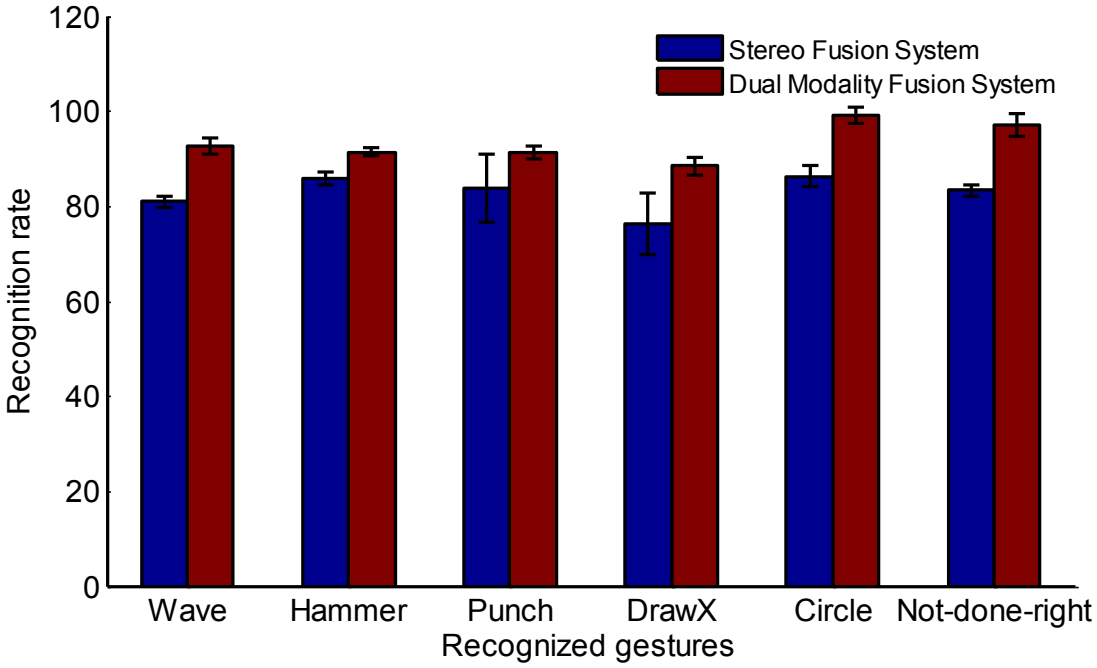

Figure 4.7. Hand gesture recognition rates (%) when using the stereo fusion system versus the dual-modality fusion system for the Microsoft Action Dataset.

As can be seen from these tables, the dual-modality fusion system outperformed the stereo fusion system. On average, the dual-modality system provided 12% higher recognition rate for the Microsoft Action Dataset and 16% higher recognition rate for the $1 Recognizer Dataset compared to the stereo system. This is attributed to the fact that the sensors in the dual-modality fusion system are of two different modalities capturing different attributes or features

of a hand gesture while in the stereo fusion system, both the left and right video images used have the same vision-based modality.

Table 4.16. Confusion matrix (% recognition rates) for the hand gestures in the $1 Recognizer Dataset when using the stereo fusion system

	A	Ca	Ch	Ci	De	Di	L	Q	Re	Ri	Sp	St	T	V	X	N
A	68	12	8	0	1	5	0	1	1	1	1	0	1	0	0	1
Ca	0	74	0	0	1	0	0	0	12	10	0	0	0	1	1	1
Ch	6	0	66	2	2	1	1	0	0	0	0	2	1	13	3	3
Ci	0	0	0	79	0	0	7	0	5	0	6	0	3	0	0	0
De	0	0	0	0	82	1	0	0	0	4	0	0	5	0	6	2
Di	0	0	0	11	1	70	4	0	6	0	4	0	3	0	0	1
L	0	0	0	10	1	5	77	0	0	3	0	0	2	0	0	2
Q	6	0	1	0	1	0	0	79	0	6	0	1	0	0	2	4
Re	0	0	0	5	0	9	1	0	72	0	0	0	8	0	5	0
Ri	1	1	0	2	2	0	4	6	0	76	0	0	0	0	5	3
Sp	0	1	1	11	0	3	0	0	0	3	79	0	0	0	0	2
St	2	0	0	3	3	7	0	0	0	0	2	73	5	0	3	2
T	0	0	0	2	2	6	0	2	5	0	0	0	78	0	4	1
V	1	2	12	3	6	0	0	0	0	0	0	0	2	73	0	1
X	0	1	1	3	7	0	0	0	0	3	0	0	0	0	83	2
N	1	2	2	1	0	1	5	3	1	4	2	1	1	3	0	73

Table 4.17. Confusion matrix (% recognition rates) for the hand gestures in the $1 Recognizer Dataset when using the dual-modality fusion system

	A	Ca	Ch	Ci	De	Di	L	Q	Re	Ri	Sp	St	T	V	X	N
A	89	3	0	0	0	0	0	0	0	0	0	0	0	2	3	3
Ca	3	90	2	0	3	0	0	0	0	0	0	0	0	0	0	2
Ch	1	2	91	0	2	0	0	0	0	0	0	0	0	4	0	0
Ci	0	0	0	87	1	3	0	0	4	0	0	0	4	0	0	1
De	0	3	4	0	86	3	0	0	0	0	0	0	0	0	3	1
Di	0	0	0	3	0	89	0	0	3	0	0	0	4	0	0	1
L	0	0	0	0	0	2	95	0	1	1	0	0	0	0	0	1
Q	0	0	0	0	0	0	0	94	0	3	0	0	1	0	1	1
Re	0	0	0	4	0	5	0	0	86	0	0	0	3	0	0	2
Ri	0	0	0	0	2	0	5	2	0	87	0	0	0	0	3	1
Sp	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	2
St	2	0	0	0	0	0	0	0	1	0	0	92	2	0	0	3
T	0	0	0	3	0	6	0	0	1	0	0	0	88	0	0	2
V	0	1	7	0	2	0	0	0	0	0	0	0	0	89	0	1
X	0	0	0	3	1	4	0	0	0	0	0	0	0	1	90	1
N	2	0	0	0	0	0	1	1	0	2	0	2	0	0	1	91

Table 4.18. provides the comparison between the two real-time systems in terms of frame rates and computational complexity. As can be seen from this table, the frame rates of the two systems were comparable with the computational complexity of the stereo system being slightly higher than the dual-modality system. The computational complexity of the stereo system is $O(\alpha m^2+L^2)$ where α denotes the number of mean shift iterations, m^2 image resolution, L the length of the warping path in the DTW algorithm.

Since the parameters of the HMM model are pre-trained, the computational complexity of the dual-modality system is basically the same as the complexity of the HMM model testing which is given by $O(RNS)$, where R denotes the number of operations to compute an observation likelihood, N the number of states in HMM, and S the number of observations [84]. It is worth pointing out that the skeleton tracking is done by a dedicated processor as part of the Kinect depth sensor [85]. As a result, the skeleton is retrieved in real-time and the skeleton image resolution has little influence on the computational complexity. Example video clips of the two systems operating in real-time can be seen at the websites [86] and [87].

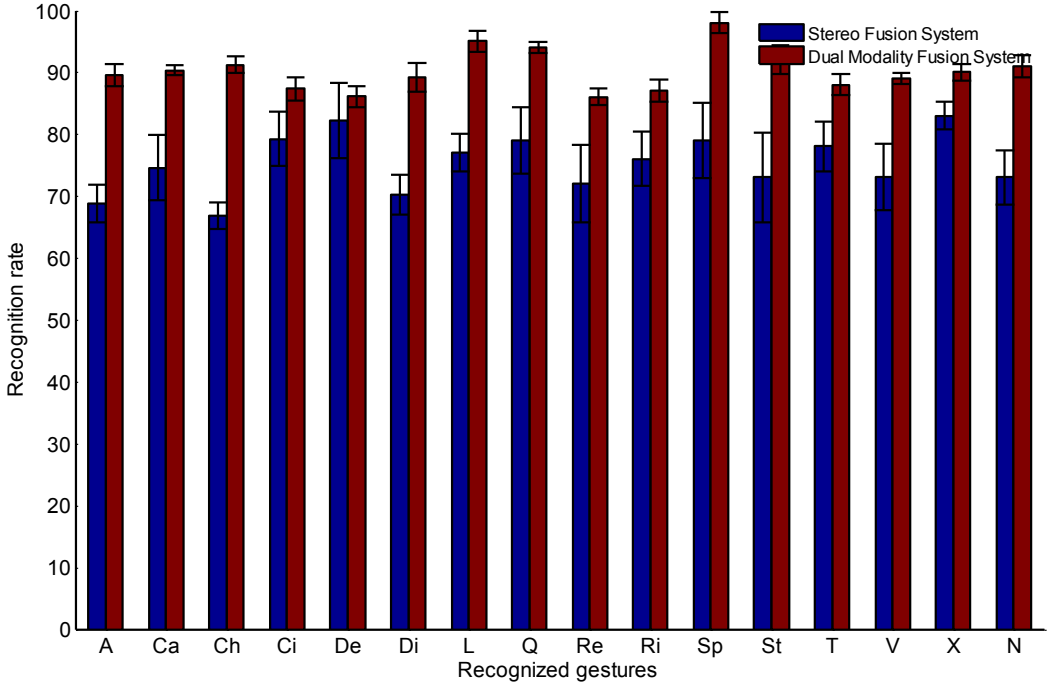

Figure 4.8. Hand gesture recognition rates (%) when using the stereo fusion system versus the dual-modality fusion system based on $1 Recognizer Dataset.

If the system is applied to embedded framework, it is necessary to discuss the performance rate with the change of the frame difference or sampling rate. Figure 4.9 gives the average gesture recognition rates (%) when using the dual-modality sensor fusion versus the stereo sensor fusion for the $1 gesture dataset. As can be seen from this figure, the frame difference in the origin is 0 which represents that the default sampling rate of the Kinect depth sensor is 30 frames per second. The recognition rate dropped significantly by using frame difference greater than 5. As illustrated in this figure, in order to keep the computational complexity low, the sampling rate was thus chosen to be 26 frames per second.

Table 4.18. Frame rates and computational complexity for the stereo fusion system versus the dual-modality fusion system

Hand gesture recognition system	Frame rates per sec	Computational complexity
Stereo Fusion	24±1.6	$O(\alpha m^2 + L^2)$
Dual-Modality Fusion	27±3.0	$O(RNS)$

4.4 MULTI-HMM CLASSIFICATION

Additional experiments were carried out to compare the performance when using the multi-HMM classification in place of the HMM classification. The code is written in C running in real-time on a PC platform with a quad core 1.7GHz processor and 4G memory. The input signals were captured with a Microsoft Kinect sensor and the inertial sensor mentioned in section 2. The inertial sensor was placed and tied to a subject's wrist.

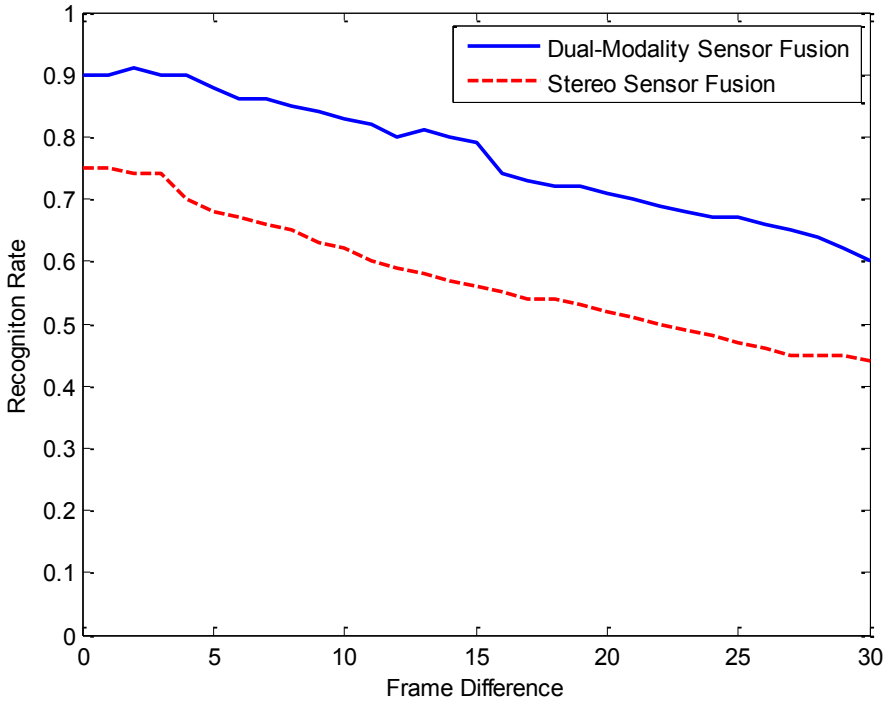

Figure 4.9. Recognition rate variation with frame difference

Ten single hand gestures of the $1Unistroke Recognizer application set [34] were considered. These gestures are illustrated in Figure 4.10 with the beginning of a gesture indicated by a solid dot.

The subjects were asked to perform the ten gestures 30 times with different speeds in different backgrounds and lighting conditions. 8-12 HMM states were used as this range of states allowed covering all the major transitions in the training sequences. The 3-axis accelerometer and the 3-axis gyroscope signals from the wireless inertial sensor and the 3-axis $\{X, Y, Z\}$ coordinates signals from the Kinect camera were captured in real-time and simultaneously to form the observation sequence $O = \{O_1, O_2, \ldots, O_T\}$. For the training of the original HMM classifier, the 9-dimensional signals were used to train one HMM classifier for each hand

gesture. While in the multi-HMM approach, three HMM classifiers were trained for each hand gesture, where each classifier was trained for 3-dimensional signals. The recognition process was repeated 10 times, each time choosing a different set of 9 training subjects. The recognition rates obtained were then averaged to remove any bias to any particular subject. For the "Not-done-right (N)" gesture category, 100 gestures were performed with 50 of them done in an incomplete way and with the other 50 done totally differently.

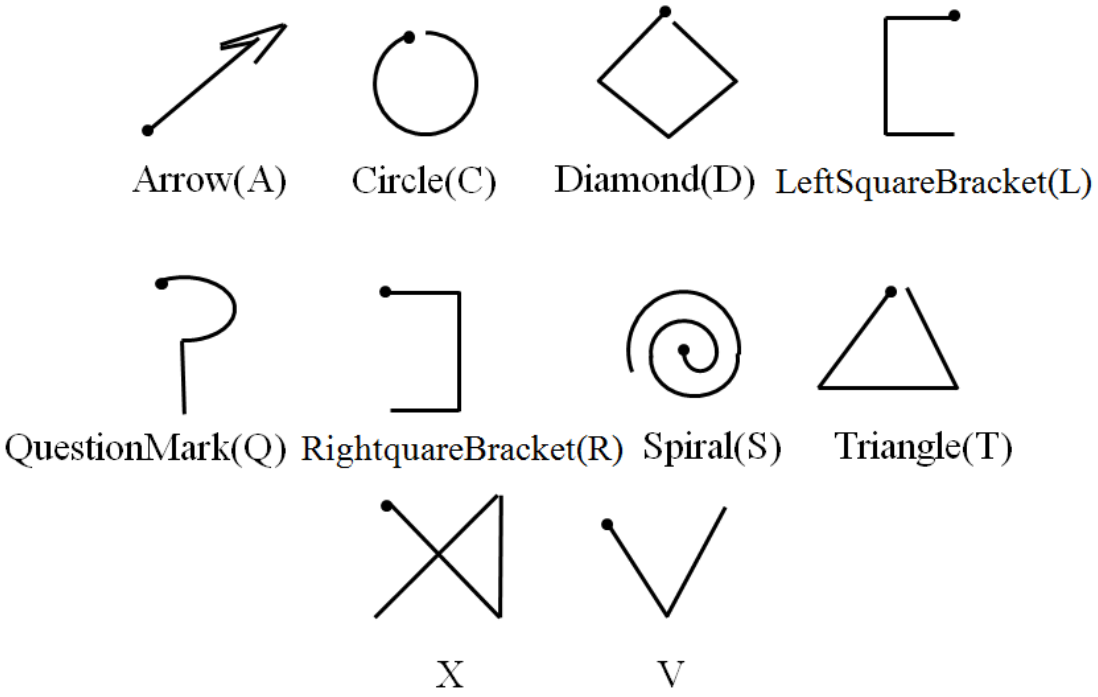

Figure 4.10. Hand gestures in the $1 Gesture Recognizer set

The recognition rates obtained are shown in the form of confusion matrices in Tables 4.18 and 4.19 for the ten studied gestures of "Arrow (A)", "Circle (C)", "Diamond (D)", "Left square bracket (L)", "Question mark (Q)", "Right square bracket (R)", "Spiral (S)", "Triangle (T)", "X", "V" and the additional class of "Not-done-right(N) ". The capital letters in the tables

represent the corresponding initials of the gestures. Table 4.18 corresponds to the situation when using the original HMM training and testing method, and Table 4.19 when using the multiple HMM training and testing method. Figure 4.11 represents the hand gesture recognition rates when using the HMM classification versus the multi-HMM classification based on the $1 Recognizer Dataset. One can see that the variances of both the HMM and the multi-HMM dataset are small. This is attributed to the fact that the dual-modality fusion system captures different attributes or features of hand gestures and thus achieves a more robust recognition rate.

As can be seen from Table 4.19, many misclassifications occurred among these gestures: “Circle”, “Diamond”, “Question mark”, “Right square bracket” and “Triangle”. This was caused due to the variance of the likelihood probabilities not being discriminatory enough to distinguish these gestures from each other.

From Table 4.20, it is seen that the multi-HMM classification led to lower misclassifications among these gestures leading to a higher overall recognition rate of 91% compared to the overall recognition rate of 84% under realistic operating conditions. Basically, this increase in the overall recognition rate was due to the enhanced discriminatory power of using the multi-HMM classification, in particular for situations involving unreliable signals from the inertial sensor or the Kinect camera. Figure 4.12 illustrates an example of normalized likelihood probability of Hand gesture QuestionMark when using the HMM classification versus the multi-HMM classification. As can be seen from this figure, the among-class difference or the variance of the recognized gesture set was improved. The multi-HMM classification exhibited more robustness.

Table 4.19. Hand gesture recognition rates (%) when using the HMM classification

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	0	2	6	7	0
C	0	82	5	0	0	0	1	10	0	0	2
D	0	8	79	1	0	0	2	7	0	0	3
L	0	7	0	90	0	0	1	2	0	0	0
Q	0	0	0	0	82	14	1	0	2	0	1
R	1	0	0	0	13	81	1	0	2	0	2
S	0	10	4	0	0	0	84	2	0	0	0
T	0	3	12	0	0	0	1	82	1	0	1
X	0	2	0	0	0	0	0	2	87	7	2
V	7	0	1	0	0	0	0	0	5	86	1
N	0	2	7	2	0	0	0	3	2	0	84

Table 4.20. Hand gesture recognition rates (%) when using the multi-HMM classification

	A	C	D	L	Q	R	S	T	X	V	N
A	88	0	0	0	0	0	0	1	4	7	0
C	0	90	5	0	0	0	2	2	0	0	1
D	0	4	86	0	0	0	1	3	2	0	4
L	0	5	0	90	0	1	1	2	0	0	1
Q	0	0	0	0	91	6	0	0	1	0	2
R	0	0	0	0	7	92	0	0	1	0	0
S	0	5	1	0	0	0	93	1	0	0	0
T	0	4	3	0	0	0	1	90	0	0	2
X	0	1	0	0	0	0	0	3	91	5	0
V	2	0	0	0	0	0	0	0	3	95	0
N	0	2	1	0	1	0	0	2	3	1	90

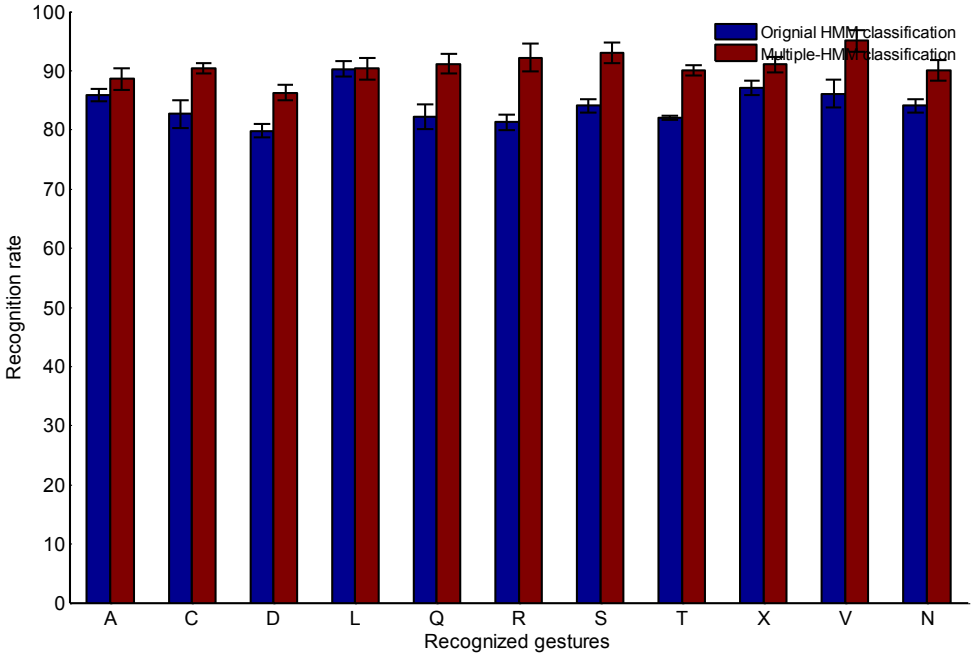

Figure 4.11. Hand gesture recognition rate (%) when using the HMM classification versus the multi-HMM classification for the $1 Recognizer Dataset.

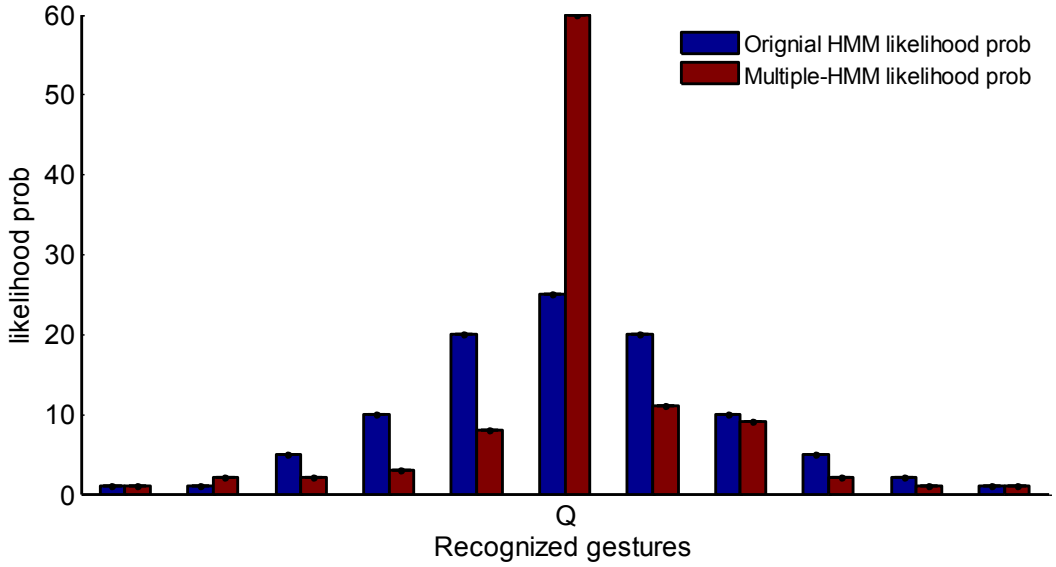

Figure 4.12. Normalized likelihood probability of the hand gesture QuestionMark (%) when using the HMM classification versus the multi-HMM classification.

4.5 SUBJECT-VARIATION STUDY

The major source of false recognitions is speed variations when different subjects perform the gestures. In a study carried out as part of this dissertation, both male and female subjects having different heights and ages were asked to perform the above gestures. The subjects performed the gestures at their own tempo or at different speeds.

It was found that when hand gestures were performed at high speeds, the fewer recorded samples of the gesture paths, see Figure 4.13, led to a lower accuracy of the gesture "Done-Right". In general, there is a trade-off between the accuracy of performing gesture trajectories and their recognition accuracies.

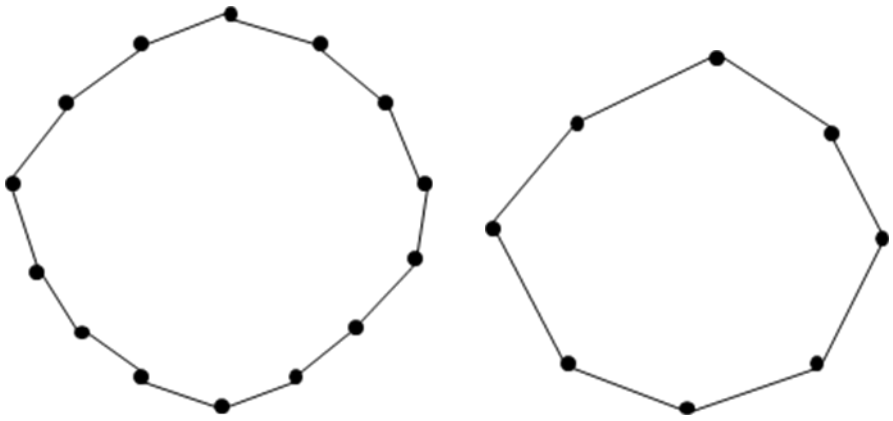

Figure 4.13. A sampled circular "Circle" gesture done at normal speed (left) and the same gesture done at fast speed (right)

Templates for different gestures, refer to Figure 4.10, were projected onto a 2D surface. Subjects wore the inertial sensor on their right wrist when the system was turned on and the gestures were initiated by smoothly moving the right hand in front of the Kinect camera at a proper distance range with the beginning of the gesture indicated by a solid dot in the figure.

Subjects were instructed to perform 10 trials. They were also told not to leave the boundary of the display screen while performing the gestures.

As shown in Figure 4.10, the circular movements such as the gestures “Circle”, “Diamond” and “Triangle”, originate from the upper middle of the screen and end in the same location. These three gestures generated the most false recognitions. This was due to the fact that when the speed of the gesture “Circle” was relatively high, the sampling of the gesture trajectory “Circle” did not include enough samples to generate accurate recognition. As a result, the gesture “Circle” was incorrectly recognized as the gesture “Diamond”. The same situation occurred for the gesture “Triangle”. The low sampling of the gesture “Triangle” caused the false recognitions of the gesture “Circle” since the corners of the gesture “Triangle” were missed. Moreover, there are four movement segments in the gesture “Diamond” while there are three in the gesture “Triangle”. If the second and the third segments of the gesture “Diamond” were not recorded properly, because of the speed in the consecutive segments being relatively high, the gesture “Diamond” was mistaken with the gesture “Triangle”.

For the same reason, the second group of gestures which generated the most false recognitions were the gesture “Question Mark” and “Right square bracket”. These gestures started from the upper left of the screen and ended in the lower middle.

The third group of gestures which generated the most false recognitions were the gesture “V” and “X”. As can be seen from Figure 4.10, there are three movement segments in the gesture “X”. When relatively tall subjects performed the gesture “X”, the skeleton of them were bigger and thus the second movement segment of the gesture fell out of the display screen. As a result, this gesture was mistakenly recognized as “V”.

For the subject-variation study, the experimental results for different subjects are shown in Tables 4.21 through 4.32, which illustrate the effect of speed variations. Note that, in general, the recognition rates using the dual modality approach outperform those using a single modality approach.

Table 4.21. Hand gesture recognition rates (%) of Subject 1 when using inertial sensor alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	5	5	0	0	5
C	0	80	0	0	0	0	5	5	0	5	5
D	5	5	75	0	0	0	0	5	0	0	10
L	0	10	0	80	0	0	0	0	0	0	10
Q	0	0	0	0	90	5	0	0	0	0	5
R	0	0	0	0	10	85	0	0	0	0	5
S	0	5	5	0	0	0	75	5	0	0	10
T	0	5	0	5	0	0	0	80	0	0	10
X	0	0	0	0	0	0	0	5	80	5	10
V	5	0	0	0	0	0	0	0	10	75	10
N	0	0	5	0	5	0	0	0	5	0	85

Table 4.22. Hand gesture recognition rates (%) of Subject 1 when using Kinect alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	5	0	5	0	10
C	0	85	0	0	0	0	5	5	0	0	5
D	0	10	80	0	0	0	0	5	0	0	5
L	0	5	0	80	0	0	0	5	0	0	10
Q	0	0	5	0	85	5	0	0	0	0	5
R	0	0	0	0	0	85	0	0	0	0	15
S	0	0	5	0	0	0	80	5	0	0	15
T	0	0	0	5	0	0	5	85	0	0	5
X	0	0	5	0	0	0	0	10	80	0	5
V	10	0	0	0	0	0	0	5	5	75	5
N	0	0	0	0	5	0	0	0	5	5	85

Table 4.23. Hand gesture recognition rates (%) of Subject 1 when using the dual-modality fusion system

	A	C	D	L	Q	R	S	T	X	V	N
A	90	0	0	0	0	0	5	5	0	0	0
C	0	85	5	0	0	0	0	5	0	0	5
D	0	10	85	0	0	0	0	5	0	0	0
L	0	5	5	90	0	0	0	0	0	0	0
Q	0	0	0	0	90	10	0	0	0	0	0
R	0	0	0	0	5	90	0	0	5	0	0
S	0	5	5	0	0	0	85	5	0	0	0
T	0	5	5	0	0	0	0	90	0	0	0
X	0	0	0	0	0	0	0	0	80	15	5
V	5	0	0	0	0	0	0	0	15	80	0
N	0	0	5	0	0	5	0	0	0	0	90

Table 4.24. Hand gesture recognition rates (%) of Subject 2 when using inertial sensor alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	0	5	0	0	10
C	0	80	5	0	0	0	0	5	0	0	10
D	0	5	85	0	0	0	0	0	0	0	10
L	0	0	0	90	0	0	0	0	0	0	10
Q	0	0	0	0	85	10	0	0	0	0	5
R	0	0	0	0	5	85	0	5	0	0	5
S	0	0	0	5	0	0	85	5	0	0	5
T	0	5	0	0	0	0	0	85	0	0	10
X	0	5	5	0	0	0	0	0	75	5	10
V	5	5	5	0	0	0	0	0	10	70	5
N	0	5	0	5	5	0	0	0	0	5	80

Table 4.25. Hand gesture recognition rates (%) of Subject 2 when using Kinect alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	5	0	0	5	5
C	0	85	5	0	0	0	0	0	0	0	10
D	0	5	85	5	0	0	0	0	0	0	5
L	0	0	5	80	0	5	0	0	0	0	10
Q	0	0	0	0	85	15	0	0	0	0	0
R	0	0	0	0	5	85	0	0	0	0	10
S	0	0	0	0	0	0	85	5	0	0	10
T	0	10	0	0	0	0	0	80	0	0	10
X	0	5	5	0	0	0	0	0	75	0	15
V	0	0	5	0	0	0	0	0	5	85	5
N	0	0	0	5	5	5	0	0	5	0	80

Table 4.26. Hand gesture recognition rates (%) of Subject 2 when using the dual-modality fusion system

	A	C	D	L	Q	R	S	T	X	V	N
A	100	0	0	0	0	0	0	0	0	0	0
C	0	85	5	0	0	0	0	5	0	0	5
D	0	5	90	0	0	0	0	5	0	0	0
L	0	0	0	95	0	0	0	0	0	0	5
Q	0	0	0	0	85	15	0	0	0	0	0
R	0	0	0	0	10	85	0	0	5	0	0
S	0	0	0	0	0	0	95	5	0	0	0
T	0	10	5	0	0	0	0	85	0	0	0
X	0	0	0	0	0	0	0	0	75	15	10
V	5	0	0	0	0	0	0	0	15	75	5
N	0	0	0	5	5	0	0	0	0	0	90

Table 4.27. Hand gesture recognition rates (%) of Subject 3 when using inertial sensor alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	0	0	0	5	10
C	5	80	5	0	0	0	0	5	0	0	5
D	0	10	85	0	0	0	0	0	0	0	5
L	0	5	0	90	0	0	0	0	0	0	5
Q	0	0	0	0	80	10	0	0	0	0	10
R	0	0	0	0	10	85	0	0	0	0	5
S	0	0	0	0	0	0	90	0	0	0	10
T	0	0	5	0	0	0	0	85	0	0	10
X	0	0	5	0	5	0	0	0	80	5	5
V	0	0	0	0	0	0	0	0	5	85	10
N	0	5	5	0	0	0	0	5	0	5	80

Table 4.28. Hand gesture recognition rates (%) of Subject 3 when using Kinect alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	0	0	10	5	0
C	0	80	5	0	0	0	0	5	0	0	10
D	0	5	85	0	0	0	0	0	0	0	10
L	0	0	0	90	0	0	0	0	0	0	10
Q	0	0	0	0	85	5	0	0	0	5	5
R	0	0	5	0	5	85	0	0	0	0	5
S	0	0	0	0	0	0	90	5	0	0	5
T	0	5	5	0	0	0	0	85	0	0	5
X	0	0	5	0	0	0	0	0	80	10	5
V	0	0	0	0	0	0	0	0	5	90	5
N	0	0	5	0	0	0	0	5	0	10	80

Table 4.29. Hand gesture recognition rates (%) of Subject 3 when using the dual-modality fusion system

	A	C	D	L	Q	R	S	T	X	V	N
A	90	0	0	0	0	0	0	0	0	5	5
C	0	85	5	0	0	0	0	10	0	0	0
D	0	15	85	0	0	0	0	0	0	0	0
L	0	5	0	95	0	0	0	0	0	0	0
Q	0	0	0	0	85	10	0	0	0	0	5
R	0	0	0	0	10	85	0	0	0	0	5
S	0	0	0	0	0	0	95	5	0	0	0
T	0	5	5	0	0	0	0	90	0	0	0
X	0	0	0	0	0	0	0	0	90	5	5
V	0	0	0	0	0	0	0	0	10	90	0
N	0	0	5	5	0	0	0	5	0	5	80

Table 4.30. Hand gesture recognition rates (%) of Subject 4 when using inertial sensor alone

	A	C	D	L	Q	R	S	T	X	V	N
A	80	0	0	0	0	0	0	5	5	0	10
C	0	80	10	0	0	0	0	0	0	5	5
D	0	5	80	0	0	0	0	5	0	0	10
L	0	5	0	90	0	0	0	0	0	0	5
Q	0	0	0	0	80	10	0	0	0	5	5
R	0	0	0	0	5	80	0	0	5	0	10
S	0	0	0	0	0	0	85	5	0	0	10
T	0	5	0	0	0	0	0	90	0	0	5
X	0	0	0	5	0	0	0	0	80	10	5
V	5	0	0	0	0	0	0	0	5	80	10
N	0	5	0	0	5	0	0	5	15	0	70

Table 4.31. Hand gesture recognition rates (%) of Subject 4 when using Kinect alone

	A	C	D	L	Q	R	S	T	X	V	N
A	85	0	0	0	0	0	0	5	5	0	5
C	0	80	5	0	0	0	0	0	0	5	10
D	0	5	85	0	0	0	0	5	0	0	5
L	0	5	0	85	0	0	0	0	0	5	5
Q	0	0	0	0	85	5	0	0	0	0	10
R	0	0	0	0	10	80	0	0	5	0	5
S	0	0	0	0	0	0	80	15	0	0	5
T	0	0	0	0	0	0	0	90	0	0	10
X	0	0	0	5	0	0	0	0	80	5	10
V	10	0	0	0	0	0	0	0	5	80	5
N	0	5	0	5	5	0	5	5	0	0	75

Table 4.32. Hand gesture recognition rates (%) of Subject 4 when using the dual-modality fusion system

	A	C	D	L	Q	R	S	T	X	V	N
A	90	0	0	0	0	0	0	5	5	0	0
C	0	85	15	0	0	0	0	0	0	0	0
D	0	15	80	0	0	0	0	5	0	0	0
L	0	5	0	95	0	0	0	0	0	0	0
Q	0	0	0	0	80	15	0	0	0	0	5
R	0	0	0	0	15	80	0	0	5	0	0
S	0	0	0	0	0	0	95	5	0	0	0
T	0	5	5	0	0	0	0	90	0	0	0
X	0	0	0	0	0	0	0	0	85	15	0
V	5	0	0	0	0	0	0	0	15	80	0
N	0	5	5	0	5	0	0	5	5	0	75

4.6 SUMMARY

In this chapter, the results of extensive experiments carried out were reported. The results obtained indicated that more robust hand gesture recognitions were achieved by using a dual-sensor approach compared to a single sensor approach. First, it was shown that by merging the information from the left and right images of a stereo image pair, an average recognition rate of 93% for seven motional hand gestures and an average recognition rate of 92% for finger spelling hand gestures were obtained under realistic lighting conditions and in various backgrounds. A careful selection of the components of the recognition system has led to a real-time processing rate of 30 frames per second on the PC platform by using an inexpensive stereo webcam.

Second, it was shown that fusing or merging the data from two differing modality sensors, consisting of an inertial sensor and a vision depth sensor, based on the probabilistic HMM classification led to an overall recognition rate of 93% for five motional hand gestures performed at different gesture speeds and in various backgrounds. This recognition rate was higher than the situations when using each sensor individually on its own.

Third, it was shown that the utilization of the multi-HMM classification in the dual-modality sensor approach led to higher recognition rates. For the ten hand gestures in the $1Unistroke Recognizer application set, an overall recognition rate of 91% was obtained under realistic conditions which included different backgrounds and lighting conditions as well as different hand speeds. This recognition rate was 7% higher than the rate when using a single component HMM classification.

CHAPTER 5

CONCLUSION

This dissertation has covered the problem of real-time hand gesture recognition using two dual-sensor approaches, one based on a stereo camera system and the other based on a Kinect depth sensor and an inertial body sensor. The dissertation places emphasis on system building or practical deployment rather than pure theoretical development. The hand gesture datasets of Microsoft Action Dataset and $1 Gesture Recognizer Dataset were used to examine and compare the performance of the developed gesture recognition systems.

The research work discussed in this dissertation was done for the purpose of increasing the robustness of recognition compared to the standard approach of using a single sensor. The contributions of the dissertation are as follows:

1. Recognition rate improvement when using a stereo camera set-up compared to when using a single camera by fusing the information from the left and the right camera in a complementary manner.

2. The introduction of a dual-modality sensor approach consisting of a Kinect depth camera and an inertial body sensor to achieve recognition rate improvement compared to the situations when each sensor is used individually on its own.

3. Utilization of a multi-HMM classification approach to improve the outcome in the dual-modality sensor approach.

5.1 STEREO CAMERA APPROACH

It should be emphasized that the use of stereo cameras here has differed from its classical use which is for obtaining depth information, rather the information from the left or the right image is used to verify the information from the other image whereby increasing the robustness of the recognition. A robust hand detection is achieved which leads to high recognition rates for two types of hand gestures. As shown in Table 4.4 and 4.6, an average recognition rate of 93% for seven motional hand gestures and an average recognition rate of 92% for finger spelling hand gestures were obtained under realistic lighting conditions and in various backgrounds. By using an inexpensive stereo webcam, a real-time processing rate of 30 frames per second on the PC platform was achieved by a careful selection of computationally efficient components.

5.2 DUAL-MODALITY SENSOR APPROACH

In Chapter 3, a dual-modality sensor approach to hand gesture recognition based on the probabilistic HMM classification was introduced for the first time. The utilization of HMM to fuse and recognize the signals from two differing modality sensors is new. The two modality sensors of Kinect depth camera and inertial sensor are chosen here because they are of low cost and they can both cope with 3D hand gestures. It was shown that fusing or merging the data from two differing modality sensors, consisting of an inertial sensor and a vision depth sensor, led to an overall recognition rate of 93% for five motional hand gestures under realistic conditions such as different gesture speeds and backgrounds as noted in Table 4.14. This recognition rate was higher than when using each sensor individually on its own.

5.3 MULTI-HMM CLASSIFICATION IN DUAL-MODALITY APPROACH

In Chapter 3, a multi-HMM classification in place of a single HMM classification was considered to improve the outcome of the dual-modality sensor approach. It was shown that for the ten hand gestures in the $1Unistroke Recognizer application set, an overall recognition rate of 91% was obtained under realistic conditions which included different backgrounds and lighting conditions as well as different hand speeds as noted in Table 4.16. This recognition rate was 7% higher than the rate when using a single component HMM classification.

5.4 POSSIBLE FUTURE WORK

Since the detection of the start and end points of the gestures has a major impact on the recognition rates, a possible future work is to utilize Kinect 2.0, which has a finger movement detection capability, to specify the start and end points of a hand gesture. Also, for the developed solutions to transition into commercial products, it is recommended to carry out further studies related to subject specific applications by performing easy training and testing involving specific subjects who will be using such a dual-modality sensing device.

REFERENCES

[1] T. Zimmerman, J. Lanier, C. Blanchard, S. Bryson and Y. Harvill, "A hand gesture interface device," *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface*, vol. 18, no. 4, pp. 189-192, Toronto, Canada, 1987.

[2] C. Keskin, F. Kirac, Y. Kara and L. Akarun, "Real time hand pose estimation using depth sensors," *Proceedings of IEEE International Conference on Computer Vision Workshops*, pp.1228-1234, Barcelona, Spain, 2011.

[3] Z. Ren, J. Meng, J. Yuan and Z. Zhang. "Robust hand gesture recognition with kinect sensor," *Proceedings of the ACM International Conference on Multimedia*, pp.759-760, Scottsdale, AZ, 2011.

[4] M. Van Den Bergh and L.Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," *IEEE Workshop on Applications of Computer Vision*, pp.66-72, Kona, HI, 2011.

[5] S. Rodriguez, A. Picon and A. Villodas, "Robust vision-based hand tracking using single camera for ubiquitous 3D gesture interaction," *Proceedings of IEEE Symposium on 3D User Interfaces*, pp.135-136, Waltham, MA, 2010.

[6] D. Gorce, D. Fleet and N. Paragios, "Model-based 3D hand pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1793-1805, 2011.

[7] Y. Ren and F. Zhang, "Hand gesture recognition based on MEB-SVM," *Proceedings of International Conference on Embedded Software and Systems*, pp. 344-349, Hangzhou, China, 2009.

[8] J. Triesch,and C. von der Malsburg, "Robust classification of hand postures against complex backgrounds," *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp.170-175, Killington, VT, 1996.

[9] F. Quek, and M. Zhao, "Inductive learning in hand pose recognition," *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp.78-83, Killington, VT,1996.

[10] A. Park, S. Yun, J. Kim, S. Min and K. Jung, "Real-time vision based Korean finger spelling recognition system," *International Journal of Electrical and Computer Engineering*, vol. 4, pp. 110-115, 2009.

[11] H. Murase and S. Nayar, "Visual learning and recognition of 3D objects from appearance," *International Journal of Computer Vision*, vol. 14, pp. 5-24, 1995.

[12] G. Bradski and J. Davis, "Motion segmentation and pose recognition with motion history gradients," *Proceedings of IEEE Workshop on Applications of Computer Vision*, pp. 238-244, Palm Springs, CA, 2000.

[13] D. Lee and K. Hong, "A Hand gesture recognition system based on difference image entropy," *Proceedings of the IEEE International Conference on Advanced Information Management and Service*, pp.410-413, Suwon, South Korea, 2010.

[14] Microsoft, "http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/"

[15] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7-42, 2002.

[16] L. Rabiner, "A tutorial on Hidden Markov Model and selected application in speech recognition," *Proceedings of IEEE*, vol.77, no.2, pp. 257-286, Feb.1989.

[17] A. Erol, G. Bebis, M. Nicolescu, R. Boyle and X. Twombly, "Vision-based hand pose estimation: A review," *Journal of Computer Vision and Image Understanding*, vol.108, no. 1-2, pp. 52-73, Oct. 2007.

[18] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Sysrtems, Man, and Cybernetics, Part C: Application and Reviews*, vol.37, no.3, pp. 311-324, May 2007.

[19] G. Murthy and R. Jadon, "A review of vision based hand gesture recognition," *Int. Jour. of Information Technology and Knowledge Management*, vol.2, no.2, pp. 405-410, July-Dec. 2009.

[20] Z. Zhang, Z. Wu, J. Chen and J. Wu, "Ubiquitous human body motion capture using micro-sensors," *IEEE Int. Conf. on Pervasive Computing and Communications*, pp. 1-5, Galveston, TX, Mar. 2009.

[21] L. Wang, T. Gu, H. Chen, X. Tao and J. Lu, "Real-time activity recognition in wireless body sensor networks: From simple gestures to complex activities," *IEEE Int. Conf. on Embedded and Real-Time Computing Systems and Applications*, pp. 43-52, Macau, Aug. 2010.

[22] M. Zhang and A. Sawchuk, "A customizable framework of body area sensor network for rehabilitation," *IEEE Int. Symp. on Applied Sciences in Biomedical and Communication Technologies,* pp. 1-6, Bratislava, Slovakia, Nov.2009.

[23] K. Liu and N. Kehtarnavaz, "Real-time robust vision-based hand gesture recognition using stereo images," *Jour. of Real-Time Image Processing*, Feb. 2013.

[24] S. Uchida and H. Sakoe, "A Survey of Elastic Matching Techniques for Handwritten Character Recognition," *IEICE Trans. On Information and systems*, vol. E88-D, no. 8, pp. 1781-1790, Aug. 2005.

[25] P. Djuric, M. Vemula and M. Bugallo, "Target tracking by particle filtering in binary sensor networks," *IEEE Trans. Signal Processing*, vol.56, no.6, pp. 2229-2238, June 2008.

[26] P. Pan and D. Schonfeld, "Video tracking based on sequential particle filtering on graphs," *IEEE Trans. Image Processing*, vol.20, no.6, pp. 1641-1651, June 2011.

[27] H. Lee and J. Kim"An HMM-based threshold model approach for gesture recognition," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol.21, no.10, pp.961-973, Oct. 1999.

[28] C. Chen, K. Liu and N. Kehtarnavaz, "Real time human action recognition based on depth motion maps," *Jour. Of Real-Time Image Processing,* 2013.

[29] Y. Wang, C. Yang, X. Wu and S. Xu, "Kinect based dynamic hand gesture recogntion algorithm research," *IEEE Int. Conf. on Intelligent Human- Machine Systems and Cybernetics*, pp. 274-279, Nanchang, China, Aug. 2012.

[30] O. Banos, A. Calatroni, M. Damas and H, Pomares, "Kinect=IMU? learning MIMO signal mappings to automatically translate activity recognition systems across sensor

modalities," *IEEE Int. Symp. on Wearable Computers*, pp. 92-99, Newcastle, UK, June 2012.

[31] S. Marcel, "http://www.idiap.ch/resource/gestures/"

[32] S. Marcel, "http://www.idiap.ch/resource/twohanded/"

[33] FGnet, "http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html"

[34] Y. Li, "Protractor: a fast and accurate gesture recognizer," *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 2169-2172 ,Paris, France, 2010.

[35] Opera, "http://www.opera.com/"

[36] Q. Bao, B. Bai, Q. Li, A. Smith, N. Vu and A. Chatziioannou, "Evaluation of the maximum a posteriori reconstruction on a microPet focus 220 scanner," *Proceedings of IEEE Nuclear Science Symposium Conference Record*, pp. 3605-3608, vol.5, Honululu, HI, Oct. 2007.

[37] A. Farag, A. Ali E. Elhabian and A. Farag, "Probability density estimation by linear combinations of Gaussian kernels-generalizations and algorithmic evaluation," *Processings of IEEE conference on multimedia technology*, pp. 6491-6494, Hangzhou, China, Jul. 2011.

[38] G. Yu, G. Sapiro and S. Mallat, "Solving inverse problems with piecewise linear estimators: from gaussian mixture models to structured sparsity," *IEEE Transaction on Image Processing*, vol. 21, no.5, pp.2481-2499, May 2012.

[39] S. Xie and J. Pan, "Hand detection using robust color correction and gaussian mixture model," *Proceedings of IEEE Conference on Image and Graphics*, pp.553-557, Hefei, China, Aug. 2011.

[40] E. Provenzi, L. De Carli, A. Rizzi and D. Marini, "Mathematical definition and analysis of the retinex algorithm," *Journal of Optical Society of America*, vol. 22, no. 12,pp.2613-2621, Dec. 2005.

[41] D. Comaniciu and P. Meer, "Meanshift: A robust approach toward feature space analysis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, May 2002.

[42] Y. Cheng, "Meanshift, mode seeking and clustering," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790-799, Aug. 1995.

[43] G. Bradski "Computer visionface tracking for use in a perceptual user interface," *Intel Technology Journal*, Q2, 1998.

[44] S. Sooksatra and T. Kondo,"CAMSHIFT-based algorithm for multiple object tracking," *Proceedings of International Conference on Computer and Infomation Technology*, vol. 209, pp. 301-310, 2013.

[45] C. Barber, D. Dobkin and H. Huhdanpaa, "The quickhull algorithm for convex hull," *ACM Transaction on Mathematical Software*, vol. 22, no. 5, pp. 469-483, Dec. 1996.

[46] E. Keogh and M. Pazzani, "Scaling up dynamic time warping to massive datasets," *Processings of European Conference on Principle of Data Mining and Knowledge Discovery*, pp.1-11, vol. 1704, Prague, Czech Republic, 1999.

[47] C. Wan and L. Liu, "Research and improvement on embeded system application of DTW-based speech recognition," *Proceedings of International Conference on Anti-conterfeiting, Security and Identification*, pp.401-404, Guiyang, China, Aug. 2008.

[48] J. Zhang and B. Qin, "DTW speech recognition algorithm of optimization template matching," *World Automation Congress*, pp. 1-4, Puerto Vallarta, Mexico, Jun. 2012.

[49] C. Faloutsos, M. Ranganathan and Y. Manolopoulos, "Fast subsequence matching in time-series database," *Proceedings of ACM Sigmod international conference on management of data*, vol. 23, no. 2, pp. 419-429, New York, Jun. 1994.

[50] J, Han and M. Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann, Jul. 2011.

[51] D. Lemire, "Faster retrieval with two-pass dynamic-time-warping lower bound," *Pattern Recognition*, vol.42, no. 9, pp. 2169-2180, Sep. 2009.

[52] M. Muller, *Information retrieval for music and motion*, Springer, Sep. 2007.

[53] B. Yi and C. Faloutsos, "Fast time sequence indexing for arbitrary Lp Norms," *Proceedings of International conference on very large data based*, pp. 385-394, Cairo, Egypt, Sep. 2000.

[54] T. Kinjo and K. Funaki, "On HMM speech recognition based on complex speech analysis," *Proceedings of IECON Annual Conference on IEEE Industrial Electronics*, pp. 3477-3480, Paris, France, Nov. 2006.

[55] J. Alon, V. Athitsos and S. Sclaroff, "Offline character recognition using online character," *Proceedings of International Conference on Document Analysis and Recognition*, vol. 2, pp. 839-843, Seoul, Korea, Sep. 2005.

[56] O. Rashid, A. Al-Hamadi and B. Michaelis, "A framework for the integration of gesture and posture recognition using HMM and SVM," *Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol.4, pp.572-577, Shanghai, China, Nov. 2009.

[57] Y. Gaus and F. Wong, "Hidden Markov model-based gesture recognition with overlapping hand-head/head-hand estimated using Kalman filtering," *Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems*, pp.262-267, Kota Kinabalu, Malaysia, Feb. 2012.

[58] J. Appenrodt, S. Handrich, A. Al-Hamadi and B. Michaelis, "Multi stereo camera data fusion for fingertip detection in gesture recognition systems," *Proceedings of IEEE International Conference on soft computing and pattern recognition*, pp. 35-40, Paris, France, Dec. 2010.

[59] X. Li and K. Hong, "Korean chess game implementation by hand gesture recognition using stereo camera," *Proceedings of IEEE International Conference on computing technology and information management*, pp.741-744, Seoul, Korea, Apr. 2012.

[60] H. Hongo, M. Ohya, M. Yasumoto and Y. Niwa, "Focus of attention for face and hand gesture recognition using multiple cameras," *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp.151-161, Grenoble, France, Mar. 2000.

[61] S. Omachi and M. Omachi, "Fast Template Matching With Polynomials," *IEEE Transaction on Image Processing*, vol. 16, no. 8, pp. 2139-2149, Aug. 2007.

[62] M. Rahman, J. Ren and N. Kehtarnavaz, "Real-time implementation of robust face detection on mobile platforms," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1353-1356, Taipei, Taiwan, 2009.

[63] K. Liu, Q. Du, H. Yang and B. Ma, "Optical flow and principal component analysis-based motion detection in outdoor videos," *EURASIP Journal on Advances in Signal Processing*, 680623, 2010.

[64] A. Bruhn, J. Weichert, and C. Schnorr, "Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211-231, 2005.

[65] P. Heckbert, *A Seed Fill Algorithm from graghics gems*, Academic Press, New York, 1990.

[66] M. Rahman, N. Kehtarnavaz, and J. Ren, "A hybrid face detection approach for real-time deployment on mobile devices," *Proceedings of IEEE Conference on Image Processing*, pp. 3233-3236, Cairo, Egypt, 2009.

[67] P. Perez, A. Blake, and M. Gangnet, "Jetstream: Probabilistic contour extraction with particles," *Proceedings of International Conference on Computer Vision*, vol.2, pp. 524–531, Vancouver, Canada, 2001.

[68] P. Soille, *Morphological Image Analysis: Principles and Applications,* Springer Verlag, New York, 2003.

[69] L. Liu, S. Zhang, Y. Zhang and X. Ye, "Human contour extraction using level set," *Proceedings of IEEE International Conference on Computer and Information Technology*, pp. 608-613, Shanghai, China, 2005.

[70] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321-331, 1998.

[71] S. Mahotra, C. Patllola, and N. Kehtarnavaz, "Real-time computation of disparity for hand-pair gesture recognition using video stereo images," *Journal of Real-Time Image Processing*, pp. 1-10, 2012.

[72] L. Wang, M. Liao, M. Gong, R. Yang and D. Nister, "High-quality real-time stereo using adaptive cost aggregation and dynamic programming," *Proceedings of IEEE International Symposium on 3D Data Processing, Visualization, and Transmission,* pp.798-805, Chapel Hill, NC, 2006.

[73] P. Senin, *Dynamic Time Warping Algorithm Review*, Information and Computer Science Department, University of Hawaii at Manoa, Tech. Rep., 2008.

[74] X. Li, J. An, J. Min and K. Hong, "Hand gesture recognition by stereo camera using the thinning method," *Proceedings of IEEE International Conference on Multimedia Technology*, pp.3077-3080, Hangzhou, China, 2011.

[75] D. Snyder, A. Hammoud, and R. White, "Image recovery from data acquired with a charge-coupled-device camera," *Journal of the optical society of America*, vol. 10, no. 5, pp. 1014-1023, 1993.

[76] T. Oskiper, H. Chiu, Z. Zhu, S. Samarasekera and R. Kumar, "Multi-modal sensor fusion algorithm for ubiquitous infrastructure-free localization in vision-impaired environments," *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pp.1513-1519, Taipei, China, Oct. 2010.

[77] C. Simaili, M. El-Najjar and F. Charpillet, "Multi-sensor fusion method using dynamic bayesian network for precise vehicle localization and road matching," *Proceedings of IEEE International Conference on tools with Artificial Intelligence*, vol.1, pp.146-151, Patras, Greece, Oct. 2007.

[78] D. Kisku, P. Gupta, H. Mehrotra and J. Sing, "Multimodal belief fusion for face and ear biometrics," *Intelligent Information Management*, vol. 1, no. 3, pp.166-171, 2009.

[79] D. Tran, D. Phung, H. Bui and S. Venkatesh, "Markov models for activity recognition using pervasive multi-modal sensors," *Proceedings of IEEE International Conference on Intelligent sensors, Sensor Networks and Information Processing*, pp.331-336, Melbourne, Australia, Dec. 2005.

[80] C. Liu and H. Fujisawa, "Classification and learning methods for character recognition: advances and remaining problems," *Machine Learning in Document Analysis and Recognition Studies in Computational Intelligence*, vol. 90, pp. 139-161, 2008.

[81] S. Connell, "Online handwriting recognition using multiple pattern class models," *Ph.D. dissertation*, Department of Computer Science, Michigan State University, May 2000.

[82] Y. Yao, P. Frasconi and M. Pontil, "Fingerprint classification with combinations of support vector machines," *Audio-and video-based biometric person authentication lectures notes in computer science*, vol. 2091, pp. 253–258, 2001.

[83] N. Karam and W. Campbell, "A multiple-class MLLR kernel for SVM speaker recognition," *Proceedings of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pp. 4117–4120, Las Vegas, NV, Mar. 2008.

[84] M. Johnson, "Capacity and complexity of HMM duration modeling techniques," *IEEE Signal Processing Letters*, vol.12, no. 5, pp. 407-410, 2005.

[85] R. Wang, *Augmented Reality with Kinect*, Packt Publishing, Birmingham, UK, 2013.

[86] Signal and Image Processing Laboratory, University of Texas at Dallas, "http://www.youtube.com/watch?v=jYg7U2UYeZo"

[87] Signal and Image Processing Laboratory, University of Texas at Dallas, "http://youtu.be/GSQrExl81mo"

[88] K. Liu, N. Kehtarnavaz and M. Carlsohn "Comparison of two real-time hand gesture recognition systems involving stereo cameras, depth camera, and inertial sensor," *Proceedings of SPIE Conference on Real-Time Image and Video Processing,* vol. 9139, May. 2014.

[89] K. Liu, C. Chen, R. Jafari and N. Kehtarnavaz, "Fusion of inertial and depth sensor data for robust hand gesture recognition," *IEEE Sensors Journal*, vol. 14, no. 6, pp. 1898-1903, Feb. 2014

[90] K. Liu, C. Chen, R. Jafari and N. Kehtarnavaz, "Multi-HMM classification for hand gesture recognition using two differing modality sensors," *Proceedings of IEEE Dallas Circuits and System Conference*, pp. 1-4, Dallas, TX, Oct. 2014.

[91] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action Recognition from Depth Sequences Using Depth Motion Maps-based Local Binary Patterns," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1092-1099, Waikoloa Beach, HI, January, 2015.

[92] C. Chen, K. Liu, R. Jafari, and N. Kehtarnavaz, "Home-based Senior Fitness Test Measurement System Using Collaborative Inertial and Depth Sensors," *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4135-4138, Chicago, IL, August 2014.

[93] C. Chen, N. Kehtarnavaz, and R. Jafari, "A Medication Adherence Monitoring System for Pill Bottles Based on a Wearable Inertial Sensor," *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4983-4986, Chicago, IL, August 2014.

[94] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-Time Human Action Recognition Based on Depth Motion Maps," *Journal of Real-Time Image Processing*, 2013.

[95] C. Chen, W. Li, E. W. Tramel, M. Cui, S. Prasad, and J. E. Fowler, "Spectral-Spatial Preprocessing Using Multihypothesis Prediction for Noise-Robust Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1047-1059, April 2014.

[96] C. Chen, W. Li, E. W. Tramel, and J. E. Fowler, "Reconstruction of Hyperspectral Imagery from Random Projections Using Multihypothesis Prediction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 365-374, January 2014.

[97] K. Liu, B. Ma, Q. Du, and G. Chen, "Fast motion detection from airborne videos using graphics processing unit," *Journal of Applied Remote Sensing*, 0001;6(1):061505-1-061505-14. doi:10.1117/1.JRS.6.061505.

[98] H. Yang, Q. Du, and B. Ma, "Decision fusion on supervised and unsupervised classifiers for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 875-879, Oct. 2010.

[99] H. Yang, B. Ma, Q. Du, and C. Yang, "Improving urban land use and land cover classification from high-spatial-resolution hyperspectral imagery using contextual information," *Journal of Applied Remote Sensing*, Aug. 2010.

[100] Q. Du, W. Wei, B. Ma, and N. H. Younan, "Hyperspectral image compression and target detection using nonlinear principal component analysis," *Proceedings of SPIE Annual Conference on Satellite Data Compression, Communication, and Processing*, San Diego, CA, Aug. 2013.

[101] B. Ma and Q. Du, "Hyperspectral targe detection with sparseness constraint," *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, Melbourne, Australia, July 2013.

[102] Q. Du, B. Ma, and N. Raksuntorn, "Improving the performance of sparse unmixing," *Proceedings of 5th IEEE GRSS Workshop on Hyperspectral and Signal Processing: Evolution in Remote Sensing*, Gainesville, FL, June 2013.

[103] B. Ma and Q. Du, "Improvement of background characterization for hyperspectral target detection," *Proceedings of IEEE 4th IEEE GRSS Workshop on Hyperspectral and Signal Processing: Evolution in Remote Sensing*, Shanghai, China, June 2012.

[104] Q. Du, J. E. Fowler, and B. Ma, "Random-projection-based dimensionality reduction and decision fusion for hyperspectral target detection," *Proceedings of IEEE Geoscience and Remote Sensing Symposium*, pp.1790-1793, July 2011.

[105] H. Yang, Q. Du, and B. Ma, "Weighted decision fusion for supervised and unsupervised hyperspectral image classification," *Proceedings of IEEE Geoscience and Remote Sensing Symposium*, Honolulu, Hawaii, Aug. 2010.

[106] K. Liu, H. Yang, B. Ma, and Q. Du, "A joint optical flow and principal component analysis approach for motion detection," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, Mar. 2010.

[107] H. Yang, B. Ma, and Q. Du, "Decision fusion for supervised and unsupervised hyperspectral image classification," *Proceedings of IEEE Geoscience and Remote Sensing Symposium*, Cape Town, South Africa, Jul. 2009.

[108] H. Yang, B. Ma, Q. Du, and L. Zhang, "Comparison of spectral-spatial classification for urban hyperspectral imagery with high resolution", *Proceedings of Joint Urban Remote Sensing Event*, Shanghai, China, May 2009.

[109] H. Su, Y. Sheng, P. Du, and K. Liu, "Adaptive affinity propagation with spectral angle mapper for semi-supervised hyperspectral band selection", *Applied Optics*, vol. 51, no. 14, pp. 2656-2663, 2012.

[110] C. Chen, W. Li, H. Su, and K. Liu, "Spectral-spatial classification of hyperspectral image based on kernel extreme learning machine", *Remote Sensing*, vol. 6, no. 6, pp. 5795-5814, 2014.

[111] H. Su, B. Yong, P. Du, H. Liu, C. Chen and K. Liu, "Dynamic classifier selection using spectral-spatial information for hyperspectal image classification", *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 085095, 2014.

[112] W. Li, K. Liu and H. Su, "Wavelet-based nearest-regularized subspace for noise-robust hyperspectral image classification", *Journal of Applied Remote Sensing*, vol. 8, no. 1, pp. 083665, 2014.

[113] H. Su, Y. Sheng, P. Du, C. Chen and K. Liu, "Hyperspectral image classification based on volumetric texture and dimensionality reduction", *Frontiers of Earth Science*, pp. 1-12, 2014.

[114] K. Liu, Y. Cheng, P. Li and T. Singh, "Source localization on two-dimensional grid," *Proceedings of IEEE Global Telecommunications Conference*, pp.1-5, 2011.

[115] J. Hu, B. J. Schroeder, and N. M. Rouphail, "Rationale for incorporating queue discharge flow into highway capacity manual procedure for analysis of freeway facilities," *Transportation research record: journal of the transportation research board*, vol. 2286, no. 1, pp. 76-83, 2012.

[116] J. Hu, and A. E. Parkany, "Transit signal priority with connected vehicle technology," *Transportation research*, vol. 38, 2014.

[117] J. Hu, B. J. Schroeder, and N. M. Rouphail, "A rationale for incorporating queue discharge flow into the HCM freeway facilities analysis procedure 2," *Transportation research*, vol. 38, 2012.

[118] Y. J. Xia, J. Hu, and M. Fontaine, "A cyber-ITS framework for massive traffic data analysis using cyber infrastructure," *The Scientific World Journal*, vol. 2013, 2013.

[119] S. X. Wei, L. Q. Ge, W. Yu, G. S. Chen, K. Pham, E. Blasch, D. Shen and C. Lu, "Simulation study of unmanned aerial vehicle communication networks addressing bandwidth disruptions," *SPIE Defense+ Security*, 90850O, 2014.

[120] W. Yu, S. X. Wei, G. B. Xu, G. S. Chen, K. Pham, E. Blasch, C. Lu, "On effectiveness of routing algorithms for satellite communication networks," *SPIE Defense, Security and Sensing*, 87390Q, 2013.

[121] W. Yu, Z. J. Chen, G. B. Xu, S. X. Wei and N. Ekedebe, "A threat monitoring system for smart mobiles in enterprise networks," *Proceedings of the ACM 2013 Research in Adaptive and covergent systems*, pp. 300-305, 2013.

[122] W. Yu, S. X. Wei, G. H. Ma, X. W. Fu and N. Zhang, "On effective localization attacks against internet threat monitors," *Proceedings of the IEEE 2013 Conference on Communications*, pp. 2011-2015, 2013.

VITA

Kui Liu, a research scientist in IFT, received the B.E. degree in electrical engineering from Nanchang University, Nanchang, China, in 2005, and the M.S. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2011. Since summer 2011, he has been a PhD student in the Department of Electrical Engineering, University of Texas at Dallas, and a member of the Signal and Image Processing Laboratory. His current research interests include motion detection, video/image processing, High performance computing, 3-D computer vision and machine learning.