Thesis for the Degree of Doctor of Philosophy

Human Pose and Activity Recognition from Stereo Images Using Probabilistic Parametric Inference

Nguyen Duc Thang

Department of Computer Engineering

Graduate School

Kyung Hee University

Seoul, Korea

August, 2011

Human Pose and Activity Recognition from Stereo Images Using Probabilistic Parametric Inference

Nguyen Duc Thang

Department of Computer Engineering

Graduate School

Kyung Hee University

Seoul, Korea

August, 2011

Human Pose and Activity Recognition from Stereo Images Using Probabilistic Parametric Inference

by

Nguyen Duc Thang

Advised by

Professor Young-Koo Lee

Submitted to the Department of Computer Engineering
and the Faculty of the Graduate School of
Kyung Hee University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Dissertation Committee:

Professor Sungyoung Lee, Ph.D. . . . . .

Professor Tae-Seong Kim, Ph.D. . . . . .

Professor Dong Han Kim, Ph.D. . . . . .

Professor Brian J. d'Auriol, Ph.D. . . . . .

Professor Young-Koo Lee, Ph.D. . . . . .

Human Pose and Activity Recognition from Stereo Images Using Probabilistic Parametric Inference

by

Nguyen Duc Thang

Submitted to the Department of Computer Engineering
on July 8, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Human pose and activity recognition has been emerged to play critical roles in numerous areas including entertainment, robotics, surveillance, etc. Here, human pose and activity recognition refers to the task of recovering the poses of a tracked subject and identifying human activities from sequential recovered poses. Usually, human poses and activities recognized over a short duration of time provide inputs to control external devices such as computers and games. Meanwhile, a long-term human pose and activity recognition adapts to proactive computing, human health-care, and discovering human lifestyles. In order to make an approach of human pose and activity recognition to be widely used, the convenience to users, the simplicity in installation, and the reasonable prices for equipment are the main factors to be considered. However, the conventional work of capturing human motion using optical markers with multiple cameras cannot totally satisfy these requirements, leading to the absence of human pose and activity recognition systems in daily applications.

Recovering human body poses and recognizing human activities from images obtained by a monocular camera may be an option. However when taking a 2-D picture of a scene with a monocular camera, we loose depth information. The appearance of a person in a 2-D image might pose many possible configurations in 3-D, that affects the results of estimating human body poses and of distinguishing alternative human activities in 3-D. In this thesis, another solution is concerned with the uses of a stereo camera: a stereo camera is a single camera consisting of two lenses to synchronously capture two images with a slight difference in the view angle from which the 3-D information of a scene can be derived to overcome the limitations of the monocular image-based approach.

The thesis demonstrates an approach of how to recover 3-D human body poses from stereo images captured by a stereo camera and an application of this approach to recognize human activities with the joint angles derived from the recovered body poses. Probabilistic parametric registration with hidden variables is applied to formulate the pose estimation approach within an efficient and generalized framework. With a pair of stereo images captured by a stereo camera, first the 3-D information (i.e., 3-D data) of a human subject is computed. Separately the human body is modeled in 3-D with a set of connected ellipsoids and their joints: the joint is parameterized with kinematic

angles. Then the 3-D body model and 3-D data are co-registered with the devised algorithm that works in two steps: the first step assigns the body part labels to each point of the 3-D data; the second step computes the kinematic angles to fit the 3-D human model to the labeled 3-D data. The co-registration algorithm is iterated until it converges to a stable 3-D body model that matches the 3-D human pose reflected in the 3-D data. The demonstrative results of recovering body poses in full 3-D from continuous video frames of various activities present an error of about 6^0–14^0 in the estimated kinematic angles. The proposed technique requires neither markers attached to the human subject nor multiple cameras: it only requires a single stereo camera.

As an application of the proposed human pose recovery technique in 3-D, an approach of how various human activities can be recognized with the body joint angles derived from the recovered body poses is presented. The features of body joints angles are utilized over the conventional binary body silhouettes and hidden Markov models are utilized to model and recognize various human activities. The experimental results show that the presented techniques outperform the conventional human activity recognition techniques.

Thesis Supervisor: Young-Koo Lee
Title: Professor

Acknowledgments

I am truly grateful to my advisor Professor Young-Koo Lee and my co-advisor Professor Tae-Seong Kim for their invaluable advice, insight, and guidance. They have advised me over the last four years since I first arrived at Korea to figure out my doctoral research topics and to complete the thesis work.

I express my sincere appreciation to Professor Sungyoung Lee, who has given me excellent supervising and guidance throughout my Ph.D. study and has provided me a terrific research environment with the Ubiquitous Computing Laboratory.

I would like to thank Professor Brian J. d'Auriol and Professor Dong Han Kim whose invaluable comments help me a lot to improve the quality of this thesis.

Many thanks to my friends in the Ubiquitous Computing Lab, especially the two senior members, Dr. Phan Tran Ho Truc and Ngo Quoc Hung, who drive me to recognize the importance of Machine Learning and to do research in a professional way. I would like to thank my friends, Dang Viet Hung, La The Vinh, and Dr. Md. Zia Uddin for their helpful comments and researching experiences and thank my roommates, Ngo Anh Vien and Hoang Huu Viet for sharing not only happiness but also difficulty in my life over several years abroad.

I am always thankful to my parents and my younger brother, whose endless love and unconditional supports have accompanied with me at every stage of my education. Without their support and encouragement, this thesis would not have been accomplished.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human Pose and Activity Recognition and Focused Research

During the last decade, automatically recognizing human poses and activities from the data acquired by sensor devices such as video sensors or attached sensors has emerged as an important research with applications in many areas. Here human pose recognition aims at recovering a human pose (i.e., a configuration of the human body) and human activity recognition (HAR) aims at recognizing a human activity (i.e., a pattern of movements of the human body) of a tracked person. Once the poses of a person changing overtime are known, the information about the body part motion is subsequently available to infer what people is doing. Thus, combining human pose recognition with a HAR engine allows us to obtain more information about human states, besides the relative position of the body limbs specified by a pose.

In general, there are two main kinds of human pose and activity recognition systems. One is a non-optical sensor based system, which uses wearable sensors. The other is an optical system (i.e., video sensor based), which uses video cameras to obtain images and applies image processing techniques to reconstruct human poses and recognize human activities from the acquired images.

In non-optical systems, the wearable sensors are attached to an exoskeleton or a suit around the human body to measure the motion of separated body limbs. The motion information is sent back to a computer, commonly throughout wireless connections, to recover whole human body poses

and to provide classifying features to distinguish human activities. Different kinds of wearable sensors have been concerned with this regard including a gyroscope to measure angular velocity or an accelerometer to measure acceleration of human body parts. So far, various commercial products to capture human motion using wearable sensors have been developed. For instance, MVN-Inertial motion capture was introduced by Xsens [5] and Gypsy by Meta motion [2].

Conventional optical systems to acquire human motion commonly use markers. Basically, the users are required to wear optical markers, so that the cameras can locate the position of the human body parts where the markers are attached. To avoid the effects of occlusion, additional cameras are installed at different locations. The number of the cameras might be up to several hundreds to make sure the full coverage around the human subject. In this system, the kinematic parameters of human poses are estimated using the relative locations of the detected markers. For instance, the kinematic angles at the knee joint are estimated based on the 3-D coordinates of the detected markers at the ankle, knee, and crotch. The main advantages of the method are fast processing speed and high accuracy. For example, capturing human body poses via VICON [4] exhibits a recording frame rate up to 240 frames-per-second that is enough to capture human activities with fast movements. Thus, such systems have been investigated mostly for pose estimation not for HAR.

Currently, markerless systems that estimate human information including poses and activities from a sequence of images without the needs of wearing markers or attached sensors are receiving more attention. Some attempts to develop marker-less systems to estimate human information from a sequence of monocular images or 2-D RGB images. Because the 3-D information of the subject is lost, the efforts to reconstruct the 3-D motion of the subject from only monocular images face difficulties with ambiguity and occlusion that lead to inaccurate results [147]. Therefore, other marker-less systems use multiple cameras to capture 3-D human motion. Through such systems, the 3-D information of the observed human subject is captured from different directional views, thereby providing better results of recovering human motion in 3-D [61, 72]. However, many

cameras may require complicated setup with extra software and hardware to support the transfers of large video data from multiple cameras over a network. Thus, there are always some tradeoffs between the flexibility of using a single camera and the ability to get the 3-D information using multiple cameras.

It is possible to obtain useful information including depth data with a stereo camera, which consists of two lenses integrated into a unified device. A stereo camera achieves depth perception in a manner similar to human eyesight. The depth information is generally reflected in a 2-D image called a depth image in which the depth information is encoded in a range of grayscale pixel values. With the flexibility in installation and convenience to users, a system to capture human pose and activity information using a stereo camera could be applicable to a wide range of applications.

An important area where the human information acquired by a stereo camera could be valuable is the field of human computer interaction (HCI). In this area, 3-D motion information is utilized to model a user by a set of joints and limbs. The motion of these joints and limbs provides efficient features to recognize human activities, which are used as inputs to control external devices such as computers and games. In conventional ways, the devices such as keyboards, joysticks, and trackballs have been the most popular techniques for acquiring the inputs from a user. However, such controllers may create a big gap between human intention and an action that a person needs to do to enter a command, requiring a user a training process to get familiar with the devices. Directly capturing human motion and using this motion to understand user's commands are therefore better options, especially for games and multimedia applications.

In healthcare applications, tracking the movements and activities of individuals may allow clinicians and family members to detect events such as dangerous falls by elderly family members, or monitor the activities of patients for diagnosis of disease. In security, a markerless system to track human motion and activity is utilized in surveillance, in which we expect an automated system to monitor people without using markers or attached sensors.

Robotics is another domain that requires human pose and activity recognition to obtain human commands. Humans are used to make communication throughout moving their hands, head, and the rest of their body. Thus, a robot, which only senses limited information from video data, cannot understand and interact with a user well. A component with its helps to exploit high level information about human poses and activities from video data plays a critical role in the developments of interactive robots.

With regards to these applications, using a stereo camera and its derived depth image is an option presented in this thesis work to develop a system to recognize both human poses and activities in 3-D. The overview of different systems and our focused research is illustrated in Fig. 1.1.

1.2 Previous Approaches

Although there are increasing interests in a single-camera based system advanced with depth-sensing ability (i.e., a stereo camera in our regard) to recognize human poses without using markers or wearable sensors, obtaining human body poses in 3-D directly from depth images is not very straightforward. Some remarkable challenges commonly arise such as the uncertainty of detecting human body parts from depth images, high dimensional kinematic parameters to model a human body, and the arbitrary appearances of human poses in 3-D.

Previously, most studies have been investigated to overcome these difficulties with the use of the *nonparametric-based* approach [27, 29, 96]. In this approach, one tries to generate a number of human pose exemplars where each is mapped to a specific depth image throughout retrieval features. Correspondingly, the retrieval features of query images are also extracted and compared against the exemplar images with their poses to find the best matching. All possible exemplars of poses can be stored in a database in advance [147]. However, this requires us a huge number of exemplars and an efficient method to organize and retrieve the poses from a database. If pose exemplars are created during human pose estimation, one needs to limit the number of created poses such as learning human movements [57]. Few studies have been attempted the *parametric-*

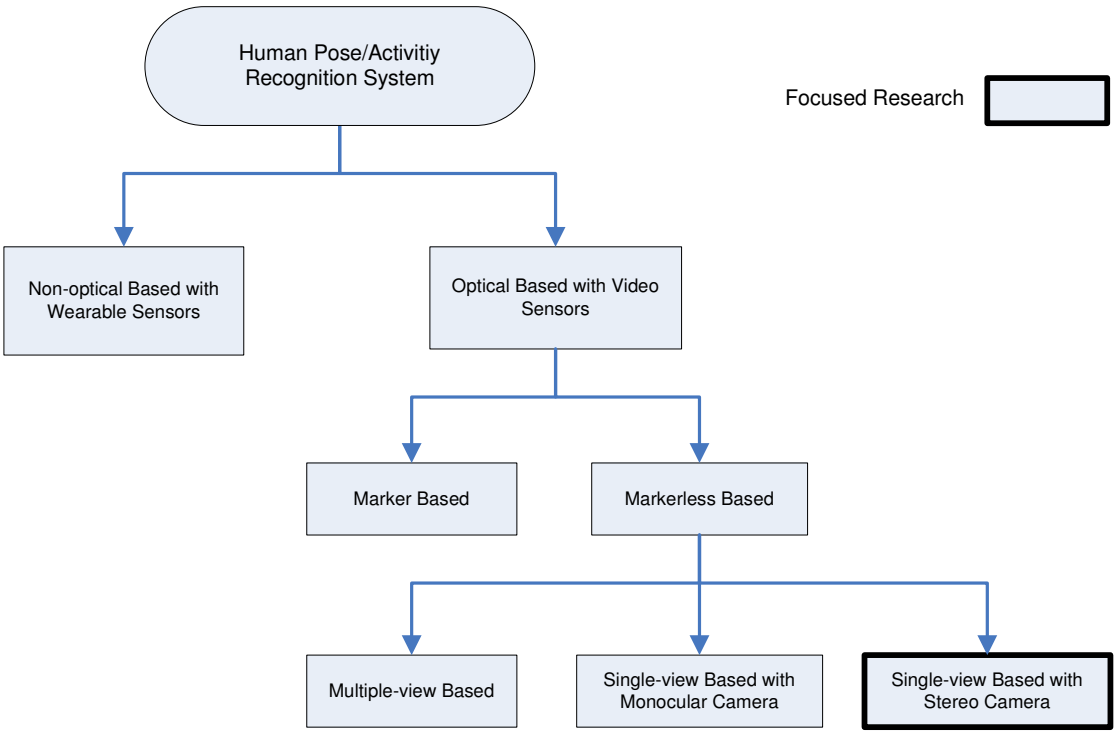


Figure 1.1: Different systems to estimate human poses and activities and our focused research.

based approach in which a parametric-based formulation is established and mathematical tools are applied for estimating human poses from stereo images without the needs of creating exemplar poses for matching.

In another aspect, previous researches of video-based HAR were concerned separately with human pose recognition. Without pose information, a video-based HAR system used parametric method with hidden Markov models (HMMs) and binary silhouette features, started from the early work of Yamato *et al.* [146]. Although binary silhouettes are commonly employed to represent a wide variety of body configurations, they also produce ambiguities by representing the same silhouette for different poses from different activities, especially for those activities that are per-

formed toward the video camera. Thus, the binary silhouettes do not seem to be a good choice to distinguish different activities.

1.3 Motivations

The ultimate goal of this thesis is to develop a system to exploit information about a person appearing in a sequence of depth images acquired by a single stereo camera. The level of information varies from the articulations of people in video to the understanding of their activities. Such discovered information will be valuable to many aforementioned applications such as human-computer interaction, health care, and surveillance.

For the pose estimation goal, as discussed in Section 1.2, most of previous studies proposed to recover human poses from depth images are based on the nonparametric approach with the requirements of creating template poses for matching. This motivates us to look for a parametric-based method to directly estimate human poses from stereo images. Parametric-based registration of a human model to video data using hidden variables (e.g., point-to-point assignments) [78, 82] might be a solution, however, how to formulate this method to estimate human poses from depths has not been developed. Thus, in this regard, we want to investigate more on the registration method with hidden variables to derive an efficient and flexible algorithm that allows us to integrate information from depths and RGB images for the task of human pose recognition. The developed technique will be valuable not only in our approaches but also in future work of recognizing human poses from different kinds of video data.

The other goal of our work is to implement an efficient HAR with the data captured by a stereo camera. However, binary silhouettes of a human body in conventional video-based HAR do not seem good enough features due to the ambiguity of 2-D information. As the human body consists of limbs connected with joints, if one can recover human poses from video images, one can form much stronger features with joint angles to improve HAR. This motivates us to look for a HAR system using joint angles of human poses recovered from depth images. With such a system, we

are able to achieve two objectives: firstly, the information about a tracked person in depth images is enriched with the understanding of human activities; Secondly, we expect an improvement in the recognition rates of the proposed HAR.

1.4 Proposed Human Pose and Activity Recognition from Stereo Images

We estimate a depth image to get 3-D information of a human subject from a pair of stereo images. We present technical challenges of recovering a 3-D human pose from a depth image as an ill-posed problem. We formulate a probabilistic registration problem of the kinematic parameters of a human body model from a depth image with the uses of hidden variables (i.e., body part labels). Our defined probabilistic framework is generalized with regards to different cues from RGB and depth images including smoothness constraints, RGB likelihoods, geodesic constraints, and reconstruction errors. Although the defined problem is complicated with the high-order priors and likelihoods of random variables, we can take advantage of inference methods that have been discovered in machine learning (see Appendix A). Here, we suggest a solution of finding an optimal pose via variational expectation maximization (VEM) to fit the defined articulated body model to depth information.

Subsequently, as an application of our technique in HAR, a sequence of kinematic angles is fed into HMMs as classifying features to distinguish different human activities of a tracked subject. We examine our proposed HAR with hundreds of stereo sequences to validate whether it is able to get better recognition rate than that of the conventional HAR approaches using body silhouette features.

1.5 Thesis Organization

We provide the thesis organization in Fig. 1.2 and the introductory of subsequent thesis chapters as follows.

- **Chapter 2** presents how to model a human body and overviews of the conventional approaches regarding the recovery of 3-D human body poses and HAR from video.
- **Chapter 3** presents our derived method to estimate human poses from stereo images.
- **Chapter 4** describes how the body poses recovered from stereo images and their joint angles can be used for HAR.
- **Chapter 5** presents the experimental results validating our proposed system to recognize human poses and activities from stereo image sequences.
- **Chapter 6** concludes the thesis with our contributions and the directions of future researches.

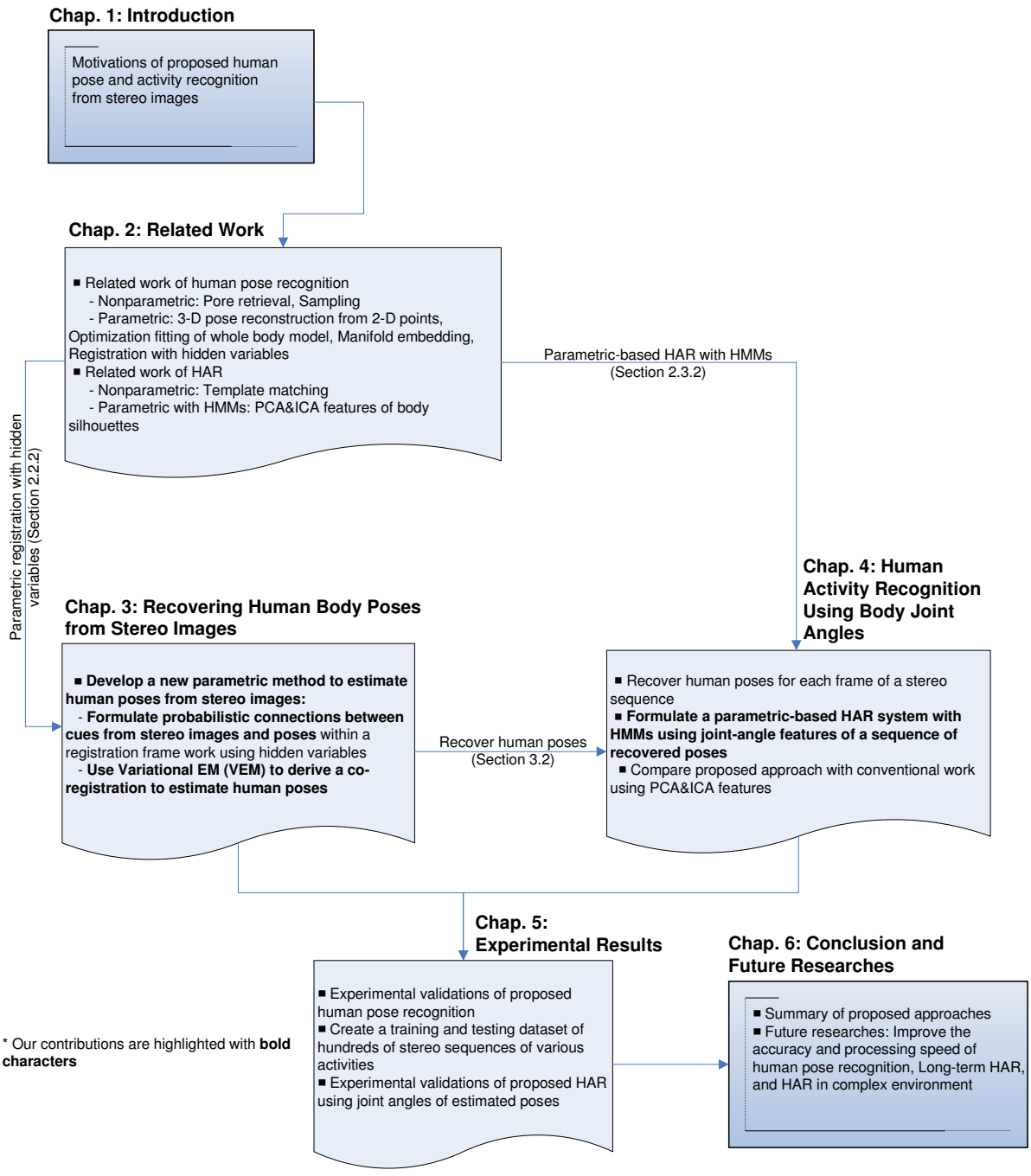


Figure 1.2: Thesis organization.

Chapter 2

Related Work

2.1 3-D Human Body Model

In general, a 3-D human model is constructed by the combination of a kinematic model to control body movements and a shape model to form a body shape.

2.1.1 Kinematic model

A kinematic model is represented by a tree consisting of body segments (i.e., a human skeletal model). Two segments are connected by a joint to allow rotation movements. As the well known result, the number of parameters necessary for a full rotation might have up to three degrees of freedom (DOF). In total, the number of kinematic parameters of the whole human body varies from 20 DOF to 60 DOF, dependent on separated studies [13, 87, 104]. Each DOF is parameterized by alternative ways including rotation matrix, Euler rotation angles, quaternion, and exponential maps. As frequently used in a human skeletal model, the shoulder is parameterized by three DOF and the elbow is parameterized by just one DOF. However, it is obviously that two DOF of the shoulder are related to the movements of the upper hand (attached to the humerus) meanwhile the other DOF of the shoulder controls the movements of the lower hand (attached to the radius).

Thus, we can reduce one DOF at the shoulder and increase one DOF at the elbow, still ensuring the movements of the body hands. Similarly, two DOF are used at every joint of of a human body. Such a configuration provides much convenience in implementation with the same number of DOF in each body joint [61, 72].

In another aspect, most kinematic models are assumed with a fixed length of body segments. To deform a human model suited with various human body shape, there have been efforts proposed to initialize a human body from images and video [21, 30]. If 3-D visual hulls of a tracked person are available, the underlying skeletal structure is able to be discovered, enabling us to obtain the length of each segment body part [21, 26, 30]. Other approaches require a manual initialization to resize a model [15] or estimate a human structure from a maker-based tracking system [120]. Fully discovering human skeleton structure and human appearances still remains challenging, requiring further investigations in future.

2.1.2 Shape model

A shape-model is designed to approximate the body shape of a tracked subject. There are two main kinds of shape models: one is a part-based model and the other is a whole-body model.

Part-based model

A part-based model represents each part of a human body by rigid objects attached to a segment of a kinematic model. Due to the rotation of each part around a joint, an instance of a human model is posed in 3-D. So far, numerous approaches have yielded success to apply part-based models for human pose estimation and human motion tracking, although such models might tolerate artifacts at body-joints where some of the model surfaces are missing. A simple implementation of part-based models was common with the use of cylinders, cones [40, 72], ellipsoids [61], and polyhedron [83]. Others modeled a human body with more complicated surfaces such as superquadric surfaces [51, 55].

Whole-body model

A whole-body model considers a single deformable surface to cover the entire shape of a human body. Such model aims at avoiding the missing information at the body-joint in the part-based model. The commonly used representations include a mesh of polygons [11] and a soft object which is expressed by a level set function in 3-D [102, 101]. The whole-body model originates in graphic areas with its applications in animation and virtual reality. Currently, the uses of such a model have been extended to estimate both human poses and shapes from image and video [9, 10, 98]. However, the complexity of creating an entire surface of a human body and the requirements of high accuracy of input sources (e.g., 3-D laser scanner) are the concerns which need to be considered with the implementations of this model.

2.2 Related Work of Human Pose Recognition

In general, there are two main approaches of human pose recognition, namely the *nonparametric-based* approach and the *parametric-based* approach. The *nonparametric-based* approach generates a number of human pose configurations where each configuration is mapped to specific features of observations (e.g., RGB images, depth images, or 3-D data). The features of query observations are extracted and used to search for the most matching poses. Alternatively, the *parametric-based* approach predefines the human body with a set of parameters related to the locations of body joints, the kinematic rotational angles, and the sizes of body parts. Then the model is fitted to the observations of video data to recover human body poses.

2.2.1 Nonparametric-based approaches for human pose recognition

Pose retrieval

One branch of method using this approach stores a large number of human pose exemplars and their matching futures in a database [62, 63, 96, 129]. Corresponding, the features from the queries

are estimated and used for retrieving the most suitable poses from a database. Thus, feature extraction and retrieval techniques become essential elements in this regard.

For 2-D images captured by a monocular camera, the internal and external contour and the binary silhouette of a human body can be utilized as the descriptors for each 3-D pose [8, 88, 110]. For the visual hulls of a human body derived from multiple cameras with multiple directional views, directly comparisons might become intractable with regard to a huge number of 3-D points belonging to a visual hull. Thus, alternative methods have been proposed to capture just essential features of observations. The 3-D Haarlet [36] presented an efficient feature due to its simplifications in calculation and its discriminant properties in classification. Linear Discriminant Analysis (LDA) [17, 148] and Average Neighborhood Margin Maximization (ANMM) [139] were used along with Haar features to reduce the dimensions of features for matching.

For a stereo camera, a set of stored poses and their corresponding depth images are compared with a depth image derived from a stereo camera to find the best matching pose. In [147], about 100,000 human poses, presenting most appearances of the human body in 3-D, were created and stored in an exemplar database. However, with a large number of human body poses, this method requires an efficient algorithm to organize and retrieve the poses stored in the database, such as parameter sensitive hashing [106, 117, 136].

Sampling

To avoid generating all possible human poses, a limited number of generated poses are limited using extra information such as cues from images, temporal information, and motion templates learned from specific activities. With a sequence of monocular images recorded with a normal camera, a probabilistic model is designed to establish the relationship between the human poses and the cues from images like color, contours, and silhouettes. Machine learning techniques such as sampling by the Monte-Carlo method [76] were applied to find the human body pose most probabilistically compatible with the information given in the images. The convergence speed

of MCMC was ensured by decomposing the Markov chain into a series of local transitions of each portion (e.g., face or limb). However, as the depth information is lost (i.e., the 3-D object is projected into a 2-D image), there will be an ambiguity of reconstructing a 3-D human pose from a monocular image. The appearance of a human subject in an image might correspond to many possible configurations of the human pose in 3-D. Due to this limitation, most previous researches based on a monocular image concentrated only on detecting the human body parts [64, 89, 105, 107, 109, 142]. The location of body parts were found by nonparametric belief propagation algorithms [122].

Besides, the approximation inference with particle filter [40, 71, 80, 118, 119, 141] was the most common techniques when sampling the whole distribution space of high dimensional random variables (30-D$\sim$40-D space of kinematic parameters) of human poses seems infeasible. Particle filter takes into account past results of human pose estimation to determine the next samples [50]: only a limited number of human poses at the time index t that are close to the human body pose estimated at the time index t-1 were generated. The effects of smoothing the motion trajectories from the past to future into the accuracy of particle-filtered human tracking were fully evaluated in [77, 100]. The drawback of this method is that with the limited number of generated poses, the accuracy of estimating human body poses tends to be low. In the opposite case, with the increased number of generated poses, the time needed to search for an appropriate human pose gets prolonged.

2.2.2 Parametric-based approaches for human pose recognition

3-D pose reconstruction from 2-D points

In this method, the articulated human body model is reconstructed from some detected regions of the human body in monocular images using inverse reconstruction 3-D from 2-D [14, 28, 42, 75, 126]. Additionally, anatomical constraints to obtain an appropriate human body skeleton were established to reduce the ambiguity of human poses, resulting a fast reconstruction of a human

body pose [145] from a 2-D image.

Optimization fitting of whole body model

A function is established to connect information from images with kinematic variables of human poses such that an estimated pose will correspond to an optimal root of this function. Typically, the information in monocular images with different directional views is combined to reconstruct the 3-D data of a human subject. Integrating the 2-D cues from each image with the data from multiple cameras, Gupta *et al.* [56] demonstrated that their system can solve the problem of pose estimation even within self occlusion. In [72], Knossow *et al.* analyzed the properties of the extremal contours of elliptical cones, then analytically derived the non-linear expressions of contour velocities that can be further used to minimize the differences between model contours and contours extracted from binary image silhouettes. The shortcomings of these methods are shown by the fact that they work separately on a single image. The outcomes also need to be combined in an additional stage to obtain the precise 3-D model parameters.

Meanwhile, with another form of representation of 3-D data, a cloud of 3-D points, in [102], the authors modeled the human body with an isosurface, called the *soft object*. The shape of the soft object was controlled by the kinematic parameters of the human model. The least-square estimator was used to minimize the differences between the soft object and the cloud of 3-D points, consequently finding the human body pose most fitted with 3-D data. In other studies, an entire mesh of a human body was deformed to fit with 3-D data of a human body [23, 137].

Manifold embedding

Rather than directly processing on images, some algorithms assume that the 3-D data are already available with 3-D voxels. To reconstruct human body poses, the 3-D data of voxels are embedded into a higher dimensional manifold. In [124], the authors presented a method to segment the 3-D voxels into different body parts and registered each part by one quadric surface to reconstruct the

articulated human model. To segment the 3-D voxels, they mapped the voxels' coordinates into a new domain using the Laplacian Eigenmaps where they could discover the skeleton structure (1-D manifolds) of the 3-D data. Based on this skeleton structure, they could assign the 3-D data to corresponding human body parts using probabilistic registration. Some other methods like ISOMAP [31, 127], Locally Linear Embedding [111], or Multidimensional Scaling [35] are also available to recover the human skeleton structure of the 3-D voxels.

Registration with hidden variables

The registration using hidden variables is the conventional method that has been applied to find a transformation to fit a set of points to others [78, 82]. In this case, the hidden variables presented the point-to-point correspondences between two datasets. In [38, 61], authors assumed that the 3-D data were drawn from a mixture Gaussian distribution where each cluster of the distribution represented a part of a human model. The kinematic parameters were found by maximum likelihood estimation with marginal integration over hidden variables. In [22], hidden variables were introduced to identify the mesh region where each 3-D point was cast to. For noisy and partial 3-D data of stereo images, it is able to extend this method of registration by exploiting information from depths and RGB images to recover human body poses, as being presented in this thesis work.

2.3 Related Work of Human Activity Recognition

For a specific video domain, a method for HAR starts with the extraction of features from images and comparing them against the features of various activities. Thus, activity feature extraction, modeling, and recognition techniques become essential elements in this regard. Approaches for modeling and recognizing activities are separated into two subcategories: the *nonparametric-based* approach extracts key features from a frame sequence and uses these features to query the best matching from stored activity exemplars; The *parametric-based* approach models dynamics of an activity and learns the modeling parameters from training data. The evaluation of fitting a

frame sequence to alternative models specifies the activity label associated with this sequence.

2.3.1 Nonparametric-based approaches for human activity recognition

The early work of the nonparametric-based approach started with a monocular image. In [19], binary silhouettes of the human body in a 2-D sequence were extracted and aggregated into an image, namely a motion energy image (MEI). If a weight was assigned to an image with regard to its chronological order, an image resulted of aggregation was called a motion history image (MHI). Correspondingly, MEI and MHI images were then utilized for the matching of two sequential images. However, two closed sequences easily created similar MEI and MHI images, leading to the ambiguity of distinguishing different activities. The other authors segmented a body contour of a person in a single 2-D image to build a surface in 3-D space (x, y, t), correspondent to a sequence of images of a specific activity [54]. The retrieval features of a 3-D surface were extracted from geometric measurements such as areas, peaks, and curvatures. In [46, 103], authors illustrated a human motion in a lower dimensional space, but this method has been better used for analyzing the motion characteristics rather than for classifying human activities. Presenting another method to reduce dimensions of observations [7], Abdelkader *et al.* located a set of 2-D points in each frame and combined the information from all sequential frames to construct a 3-D deformable model, which was used for classification.

The drawback of the nonparametric-based approach is stated that it only obtains good results with recognizing simple and short-time activities [131]. Also, there is not much attention from research communities to utilize 3-D data because the template of a 3-D object moving over time will be aggregated in a 4-D space-time, leading to the difficulties of extracting retrieval features to characterize an activity.

2.3.2 Parametric-based approaches with HMMs for human activity recognition

HMMs are the most common video-based model of human activity that have been applied for parametric-based HAR. For instance, in [146], a binary silhouette-based HAR system was proposed to transform the time sequential silhouettes into a feature vector sequence through the binary pixel-based mesh feature extraction from every image. Then, the features were utilized to recognize several tennis actions with HMMs. In [24], a silhouette matching key frame-based approach was applied to recognize forehand and backhand strokes from tennis videos. Regarding binary silhouette-based features, Principal Component Analysis (PCA), a feature extractor based on the second-order statistics, is most commonly applied [93, 94, 132]. After applying PCA, some top PCs (i.e., eigenvectors) are chosen to produce global features representing most frequently moving parts of the human body in various activities. In [93, 94], the authors utilized PC features from binary silhouettes and optical flow-based motion features in combination with an HMM to recognize different view-invariant activities.

Recently, more advanced HAR techniques have been introduced in terms of new features and more powerful feature extraction techniques such as Independent Component Analysis (ICA) of body silhouettes [132, 133]. Although binary silhouettes are commonly employed to represent a wide variety of body configurations, they also produce ambiguities by representing the same silhouette for different poses from different activities, especially for those activities that are performed toward the video camera. Thus, the binary silhouettes do not seem to be a good choice to represent human body poses in different activities. In this regard, more efficient features exploited from the depth information should be a solution to get better results of human activity recognition.

Chapter 3

Recovering Human Body Poses from Stereo Images

In this chapter, we present a technique of estimating 3-D human body poses from a set of sequential stereo images. We developed a new algorithm based on the *parametric-based* approach to estimate human body poses directly from stereo images without using a set of temporary poses for matching. Among methods concerning this approach, our implementation is based on the model-to-data registration with the uses of hidden variables to indicate body part labels, as introduced in Section 2.2.2 of Chapter 2. The rest of this chapter is organized as follows. In Section 3.1, we describe our methodology. The main algorithm of recovering human poses from 3-D data is presented in Section 3.2 and summarized in Section 3.3.

3.1 Methodology

The step-by-step processing stage of our system is briefly described in Fig. 3.1. In the preprocessing step, we estimate the disparity between the left and right images taken by a stereo camera. The 3-D location of the observed subject is reconstructed using disparity values and represented

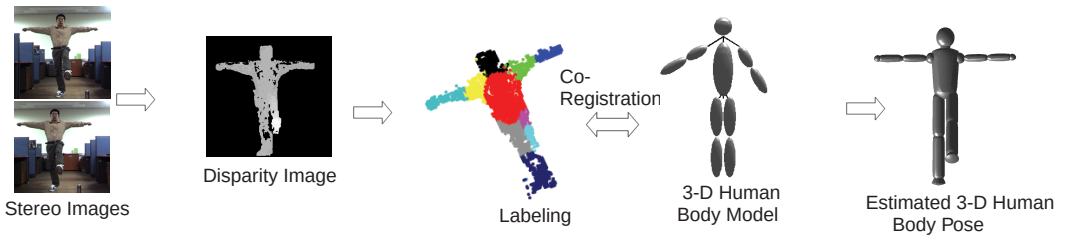


Figure 3.1: Our proposed method of estimating a 3-D human body pose from stereo images. (a) A set of stereo images. (b) Estimated disparity image. (c) Labeling the body parts of the 3-D data. (d) Fitting the 3-D model with the 3-D data. (e) Final estimated body pose.

by a cloud of points in 3-D. To fit the 3-D model to the given 3-D data, we perform co-registration with VEM in two steps: VE-step and model fitting (M-step). The VE-step assigns each point to one ellipsoid and the model fitting step fits the ellipsoids to their corresponding points. This process is iterated by minimizing the discrepancies between the model and the observation, finally recovering the correct human pose. The details of our co-registration algorithm are discussed in Section 3.2.

3.1.1 Stereo camera and stereo image processing

Stereo camera

Through several million years of human evolution, stereopsis is one of the unique functions in the human vision system, allowing depth perception: it is a process of combining two images projected to two human eyes to create the visual perception of depth. Learned from the human stereoscopic system, a stereo camera was invented to synchronously capture two images of a scene with a slight difference in the view angle from which depth information of the scene can be derived. The depth information is generally reflected in a 2-D image called a depth image

in which the depth information is encoded in a range of grayscale pixel values. Since its first commercial product in 1950s, Stereo Realist, introduced by the David White Company, there have been continuous developments of a stereo camera until now with the latest products such as a digital stereo camera, Fujifilm FinePix Real 3-D W1 [1] and a stereo webcam, Minoru 3-D [3]. Lately, 3-D movies, in which depth information is added to RGB images, have received a lot of attention with the latest success of a film, Avatar released in 2009. Watching 3-D movies and 3-D TVs with the special viewing glasses is becoming a part of our lives these days.

In this work, a stereo camera is valuable for human pose estimation. We employ the stereo camera Bumblebee 2.0 of Point Grey Research [6] to capture stereo image pairs, as shown in Fig. 3.2. Bumblebee 2.0 camera is equipped with two Sony 1/3" progressive scan CCDs, Color/BW sensors, which are able to capture an image with a resolution of 640×480 and 1024×768 and with a speed of 20$\sim$40 frame per second (FPS). The IEEE-1394a FireWire interface is used to connect a stereo camera with a computer with a bandwidth of 400Mb/s. Also, the camera is supported with integrated functions to pre-calibrate recorded images against distortion and misalignment.

Stereo computation

The computation of stereo information is the preliminary processing step necessary to recover 3-D information from a pair of stereo images. The displacements between two images are presented as a depth image containing the disparity values. With an ordinary searching technique, it exhausts $O(n^3)$ computation to obtain the complete disparity values, assuming that the size of the image is n^2 [60, 90, 114]. We use the fast stereo matching algorithm, Growing Correspondence Seeds (GCS) [25], which requires only a small fraction of the disparity space to improve speed and accuracy. The computation complexity becomes $O(kn^2)$ with $k \ll n$ compared with searching the entire disparity space at $O(n^3)$. Moreover, if the background is partially eliminated, we can reduce the searching time on the sparse regions. The approach we apply for the background modeling and removal is described in [140].

Figure 3.2: Stereo camera Bumblebee 2.0 of Point Grey Research.

Then, the depth image is sampled by a grid to reduce the number of points in the observed data and avoid extensive computation, as depicted in Fig. 3.3(b). To obtain the 3-D data, the depth value Z of each point is computed by

$$Z = \frac{fb}{d} \tag{3.1}$$

where f is the focus length, b is the base-line, and d is the disparity value. The two remaining coordinates X and Y are given by

$$X = \frac{uZ}{f}, Y = \frac{vZ}{f} \tag{3.2}$$

where u and v are the column and row index of a pixel in the depth image.

3.1.2 3-D human body model

Our 3-D human model is reconstructed by the combinations of a kinematic model using two DOF at each body joints (see Section 2.1.1) and a part-based model of ellipsoids (see Section 2.1.2). In the computation of transformation, we formulate the equation of an ellipsoid [61] in the 4-D projective space as

$$q(X) = X^T \mathbf{Q}_\vartheta^T \mathbf{S}^T \mathbf{D} \mathbf{S} \mathbf{Q}_\vartheta X - 2 = 0 \tag{3.3}$$

where $\mathbf{D} = diag[a^{-2}, b^{-2}, c^{-2}, 1]$ configures the size of the ellipsoid, $\mathbf{S}$ locates the center of the ellipsoid in the local coordinate system, $\mathbf{Q}_\vartheta$ is the skeleton-induced transformation, and $X =$

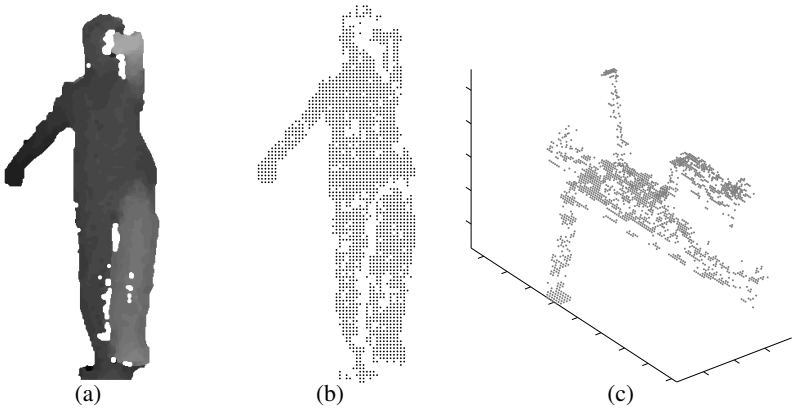


Figure 3.3: Computing the 3-D stereo data. (a) Depth image. (b) Sampling on the grid. (c) 3-D data.

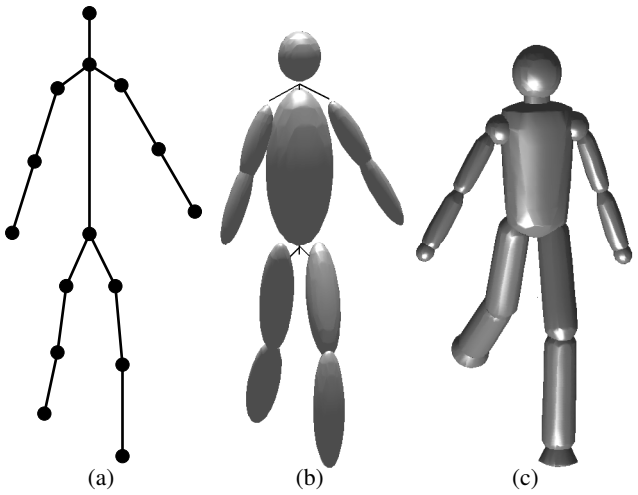


Figure 3.4: 3-D human body model. (a) Skeleton model. (b) Computation model with ellipsoids. (c) Human synthetic model with super-quadrics.

$[x, y, z, 1]^T$ is the coordinate of a 3-D point. We choose $b = a$ and $c \geq a$ to simplify the Euclidean distance computation from one point to an ellipsoid. The 4x4 transformation matrix $\mathbf{Q}_\vartheta$ is a matrix function of $\vartheta = (\vartheta_1, \vartheta_2, ..., \vartheta_n)$ where $\vartheta_1, \vartheta_2, ..., \vartheta_n$ are the n kinematic parameters that control the position of each ellipsoid in the model. $\mathbf{Q}_\vartheta$ is not only a single transformation, but it relates to a kinematic chain of transformations through each body part. The joint between two adjacent parts has up to three rotational DOF, while the transformation from the global coordinate system to the local coordinate system at the human hip requires six DOF (i.e., three rotations and three translations). We separate $\mathbf{Q}_\vartheta$ to a series of independent primitives that only depend on a single parameter,

$$\mathbf{Q}_\vartheta = \mathbf{Q}_n(\vartheta_n)\mathbf{Q}_{n-1}(\vartheta_{n-1})...\mathbf{Q}_1(\vartheta_1) \tag{3.4}$$

where $\mathbf{Q}_1(\vartheta_1), \mathbf{Q}_2(\vartheta_2), ..., \mathbf{Q}_6(\vartheta_6)$ are of six DOF of the global transformation and $\mathbf{Q}_i(\vartheta_i) = \mathbf{Tr}_i\mathbf{R}(\vartheta_i)$ with $i > 6$ is the local transformation from one coordinate system i to the other $i + 1$. $\mathbf{Tr}_i$ is the translation matrix determined by the skeleton architecture and $\mathbf{R}(\vartheta_i)$ is the rotation matrix around the $x-, y-$, or $z-$axis. We can set $\mathbf{Tr}_i$ to be the identity matrix $\mathbf{I}_{4\times4}$ if we want to add more than one DOF to a joint.

The whole body configuration is depicted in Fig. 3.4. There are 14 segments of the body, nine joints (two knees, two hips, two elbows, two shoulders, and one neck), and 24 DOF (two DOF at each joint [61] and six free transformations from the global coordinate system to the local coordinate system at the hip). Each body part may contain several ellipsoids. However, to simplify the computation, we use only one for each.

For better display and to create a synthetic human model for simulations, we also designed a model using super-quadrics as shown in Fig. 3.4(c). The equation of the super-quadric surface [37, 124] without any transformation is expressed as

$$\left(\frac{x}{a_0}\right)^2 + \left(\frac{y}{b_0}\right)^2 = \left(1 + \frac{sz}{c_0}\right)\left(1 - \left(1 - \frac{2z}{c_0}^d\right)\right)$$
$$0 \leq z \leq c_0 \tag{3.5}$$

where a_0, b_0, and c_0 determine the size of the super-quadric along the x-axis, y-axis, and z-axis, respectively.

3.1.3 Distance from one point to an ellipsoid

The distances between a set of points to an ellipsoid are used to measure the differences between the 3-D data and the model. For simplification, the function $q(X)$ defined in (3.3), which approaches zero at the ellipsoid surface and becomes larger when the point moves away from the ellipsoid, has been defined as the *algebraic distance* [102]. However, due to variation that is related to direction (e.g., with the prolate spheroid, the algebraic distance gets smaller as the point moves toward the poles), the algebraic distance cannot exactly reflect the measurement, especially for thin ellipsoids (usually representing limbs). In addition, Horaud *et al.* [61] proposed an alternative distance, the *datum distance*; however, as it requires normal vectors, it is very difficult to calculate this distance from the data gathered by a stereo camera alone.

The *Euclidean distance*, equal to the distance from one point to its nearest point in the ellipsoid surface, is rarely used because it requires solving a sixth-degree polynomial equation [58]. In this work, with the symmetric ellipsoid model, the calculation of Euclidean distance can be simplified: first of all, rather than computing Euclidean distance in the global coordinate system (x, y, z), the point $X_0(x_0, y_0, z_0)$ can be transformed to the local coordinate system (x', y', z') that holds the ellipsoid. In Fig. 3.5, let P be the plane that contains a point X_0 and the major z'-axis of the ellipsoid. The intersection between the plane P and the ellipsoid will be an ellipse. The computation of the Euclidean distance to an ellipsoid is reduced to find the distance between a point X_0 and an ellipse lying in P with only a fourth-degree polynomial equation that has an analytical solution enabling us to calculate its roots.

Moreover, the kinematic parameter $\vartheta = (\vartheta_1, \vartheta_2, ..., \vartheta_n)$ in (3.3) is updated by the gradient descent method in Section 3.2.2. Therefore, at each step, the point X_0 moves to $X_0 + dX_0$ with a small change dX_0 in the local coordinate system (x', y', z'). Corresponding, X_t, the nearest point

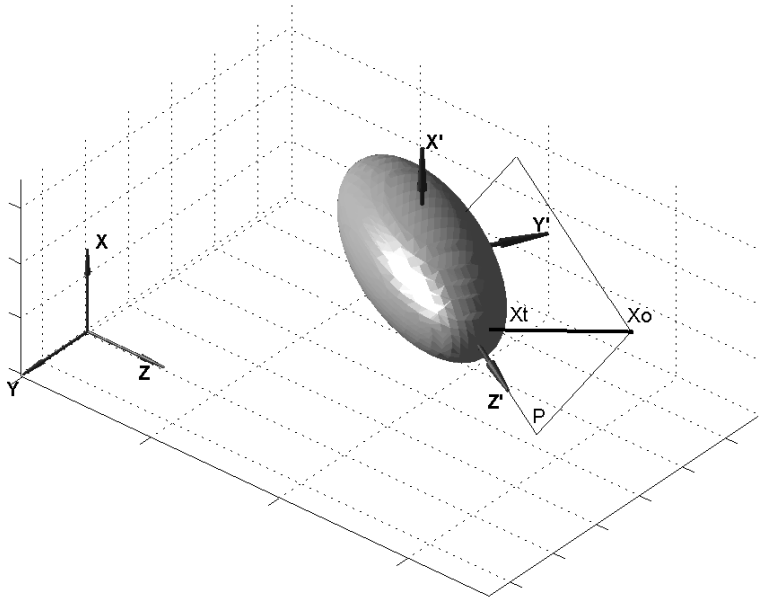


Figure 3.5: The Euclidean distance from a point to an ellipsoid.

of X_0 in the ellipsoid surface, also moves to $X_t + dX_t$, which can be calculated from X_0, dX_0, and X_t with some multiplication and addition.

The mathematical details of finding the nearest point in an ellipsoid surface to a given point are described in Appendix D.

3.2 Estimating 3-D Human Body Pose from 3-D Stereo Data

This section presents our algorithm to estimate 3-D human body pose from the 3-D stereo data. First, we establish a comprehensive conditional probabilistic distribution between the human pose specified by the kinematic parameter $\vartheta = (\vartheta_1, \vartheta_2, ..., \vartheta_n)$ and the given 3-D data and RGB image. Then, we show how to estimate the optimal kinematic parameter ϑ^* that maximizes the distribution by the VEM algorithm. The estimated parameter ϑ^* will correspond to the most suitable human pose with the given information.

3.2.1 Probabilistic relationship between the model parameters and the stereo data

We use $D = (X_1, X_2, ..., X_M)$ to denote M points of the 3-D data and I for the RGB image. Since our model is created with multiple ellipsoids, the supplementary variables are introduced to determine to which part of the body (i.e., ellipsoid) each point should belong. Let $V = (v_1, v_2, ...v_M)$ denote the body part assignments or labels of each point. The posterior probability of the label V and the model parameter ϑ given the 3-D data and RGB image is expressed by

$$P(V, \vartheta|I, D) \propto P(V)P(I|V)P(D|V)P(D|V, \vartheta). \tag{3.6}$$

The elements of (3.6) are sequentially defined in the following sections.

Smoothness energy

The smoothness prior $P(V)$ is derived in the form of the Potts model [20],

$$P(V) = \prod_{i=1}^{M} \prod_{j \subset \mathcal{N}_i} P(v_i, v_j) \tag{3.7}$$

where $\mathcal{N}_i$ is a set of neighbors of point i and $P(v_i, v_j)$ is,

$$P(v_i, v_j) = \begin{cases} e^{\gamma} & \text{if } v_i = v_j \\ 1 & \text{if } v_i \neq v_j \end{cases} \tag{3.8}$$

where γ (in our case $\gamma = 0.5$) is a real positive constant. $P(v_i, v_j)$ is used to drive the label of each point toward the same label of its neighbors. This causes the labeling results to become smooth and eliminates the outliers. The simplest way to locate the neighbors bounded by the radius d of one point is via a mask. We predefine the binary mask based on the distance d and perform an operation via the AND operator with the binary silhouette to find the neighbors of each point. We set $d = 2$ for all of our experiments.

Image likelihood

Some partial regions in the RGB image can provide extra information to identify the body components. Generally, the image likelihood term is derived as

$$P(I|V) = \prod_{i=1}^{M} \phi(I|v_i). \tag{3.9}$$

One might utilize the shape of the binary silhouettes or texture information to detect body parts. In our approach, we apply face detection to locate the head. Potential face areas are ascertained by detecting skin in the HSV color space and thermal infrared domains [34]. Some regions lying outside the binary silhouette or having unsuitable shapes (too small or appearing to be limbs) are considered outliers and removed. Estimation of the binary silhouette that relies only on background subtraction is not enough to obtain the correct result due to the effects of

(a) (b) (c)

Figure 3.6: Binary silhouette extraction. (a) Input image. (b) Background substraction. (c) Refined silhouette.

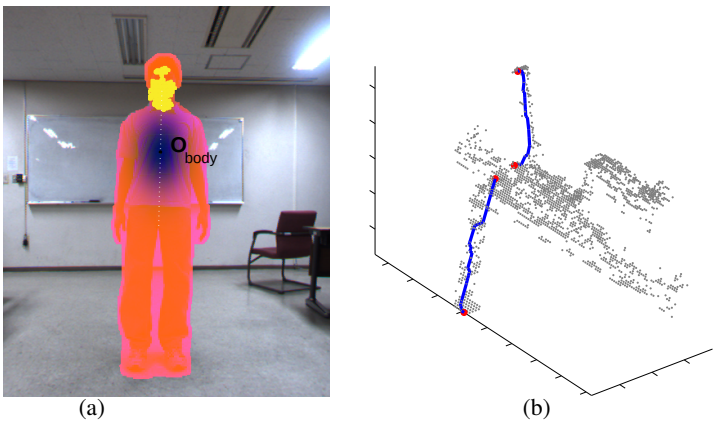
lighting conditions and shadows. As shown in Fig. 3.6, after the stereo computation, based on the estimated distance between the person and the camera, some pixels remaining outside the ranges are removed to refine the silhouette. $\phi(I|v_i = head)$ evaluating the likelihood of point i to be assigned the label *'head'* gets a value of e^c $(c = 1)$ for the pixel marked as '*faces*' and a value of one in other cases.

Together with face detection, an additional function $f(\mathbf{x}_i)$ (related to the concept of soft objects [102]) is defined to estimate the torso location. If we let the center of the body O_{body} lie at a middle point between the center of the face and the center of the silhouette. $f(\mathbf{x}_i)$ is computed in the following way:

$$f(\mathbf{x}_i) = \kappa e^{-d(\mathbf{x}_i)} \tag{3.10}$$

where $d(\mathbf{x}_i)$ is the algebraic distance from the point $\mathbf{x}_i = [x, y, 1]^T$ to the ellipse with the centroid O_{body} and κ $(\kappa = e)$ is a positive constant. In the coordinate system attached to the origin O_{body}, $d(\mathbf{x}_i) = {\mathbf{x}_i}^T \mathbf{Q}_e^T \mathbf{D}_e \mathbf{Q}_e \mathbf{x}_i - 1$ where $\mathbf{D}_e$ and $\mathbf{Q}_e$ are the 3×3 matrices that determine the shape and orientation of the ellipse. The likelihood for identifying a single point as '*torso*' is given by

$$\phi(I|v_i = torso) = \begin{cases} f(\mathbf{x}_i) & \text{if } d(\mathbf{x}_i) \leq 1 \\ 1 & \text{otherwise.} \end{cases} \tag{3.11}$$



(a) (b)

Figure 3.7: Illustration of the factors that affect label assignments. (a) Image likelihood for detecting the face and torso. (b) Geodesic distance preserved with human movements.

Pairwise geodesic relationship among 3-D points

The Euclidean distance between any two points is only preserved within a rigid object. With a non-rigid object like the human body, the Euclidean distance will be changed due to the non-linear deformations of various body parts while the object is moving. However, with regard to the geodesic distance between a pair of points in space, this distance always retains its value during the movement of a human body. The preservation of geodesic distance has been successfully applied by the ISOMAP algorithm [127] to determine the manifold of high-dimensional data in a lower dimension. Here, we attempt to represent the geodesic relationships between each point and others in our probabilistic model. Some constraints are established to restrict the probability of incorrect label assignments. Assigning the pixels into groups called cells, as illustrated in Fig. 3.8, can help us save the computational time. All of the elements belonging to the same cell receive the same geodesic constraints. The geodesic distance between two cells is approximated by the shortest path distance in a graph using Dijkstra's algorithm [43]. The compatible probability $P(D|V)$ of the 3-D data with the geodesic constraints is given by

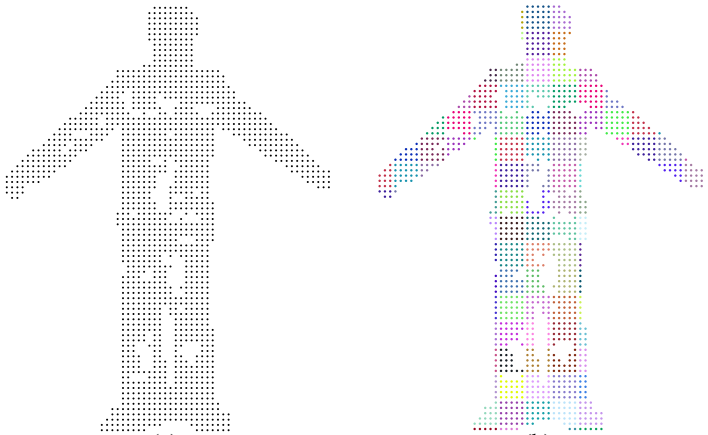


Figure 3.8: Assigning points into cells. (a) Sampling on the grid. (b) Points grouped by cells.

$$P(D|V) = \prod_{i=1}^{M} \prod_{j_c=1}^{M_c} P_{geo}(D|v_i, v_{j_c})$$
$$P_{geo}(D|v_i, v_{j_c}) = \begin{cases} e^{-\alpha} & d(v_{i_c}, v_{j_c}) < d_{min}(v_{i_c}, v_{j_c}) \\ e^{-\beta} & d(v_{i_c}, v_{j_c}) > d_{max}(v_{i_c}, v_{j_c}) \end{cases} \tag{3.12}$$

where i_c is the cell that holds pixel i, $d(v_{i_c}, v_{j_c})$ is the geodesic distance between the cell i_c and j_c, M_c is the number of cells, and α and β are two positive constants. Two values, $d_{min}(v_{i_c}, v_{j_c})$ and $d_{max}(v_{i_c}, v_{j_c})$, define the lower and upper bounds for the geodesic distance between a pair of labels. Two related labels that are too far or too close are penalized to decrease the belief in those labels. The constant values α and β are taken to be $\alpha = \beta = 0.04M_c/M_{v_{j_c}}$, limiting the maximum number of cells which can ascertain to the label of the pixel i to 4% of the total. The maximum number of cells receiving the same label v_{j_c}, $M_{v_{j_c}}$, appears in the denominator as a normalized constant to ensure that $\prod_{j_c=1}^{M_c} P_{geo}(D|v_i, v_{j_c})$, the total effect to the pixel i, is approximately invariant to the size of body parts.

Reconstruction error

To co-register the ellipsoid model with the observations, we need to minimize the differences between them. The last term accounts for the compatible probability between the model specified by ϑ and the data D consisting a set of points $X_1, X_2, ..., X_M$. Let's denote $d(X_i, \vartheta, v_i)$ as the Euclidean distance between a point $X_i(x_i, y_i, z_i)$ and an ellipsoid v_i, as we already discussed in Section 3.1.3. $P(D|V, \vartheta)$ is defined as

$$P(D|V, \vartheta) = \prod_{i=1}^{M} e^{-\frac{d^2(X_i, \vartheta, v_i)}{2\sigma^2}} \tag{3.13}$$

where σ denotes the variance (σ^2 is chosen to be 0.1 in our experiments). The distance between the point X_i and the ellipsoids is also one of the factors that decides the body segment of X_i. Hence, in a sequence of frames, the estimated model from the current frame presents a good initial model to derive the label on the next frame.

3.2.2 Estimating the model parameters

Our main goal is to find the optimal kinematic parameter ϑ^* that maximizes the posterior probability of ϑ given the data. This problem can be rewritten as

$$\vartheta^* = \text{argmax}_{\vartheta} \sum_{V} P(V, \vartheta|I, D) \tag{3.14}$$

where V is considered the latent variable in this framework. The VEM algorithm is a good choice for estimating the optimal values of the probabilistic problem with the appearances of unobserved variables (see Appendix C). By introducing the distribution $Q(V)$ over the variable V [18], the problem in (3.14), equivalent with maximizing $\sum_V \log P(V, \vartheta|I, D)$, can be decomposed into

$$\begin{aligned} &\text{argmax}_{\vartheta, Q} \sum_{V} Q(V) \log \frac{P(V, \vartheta|I, D)}{Q(V)} \\ &- \sum_{V} Q(V) \log \frac{P(V|\vartheta, I, D)}{Q(V)}. \end{aligned} \tag{3.15}$$

The VEM algorithm is an iterative procedure whose each iteration consists of the following two main steps:

i) Assuming that the current value of ϑ is $\vartheta_{\textbf{old}}$, the VE-step approximates the posterior distribution $P(V|\vartheta, I, D)$ as

$$Q(V) \propto P(V|\vartheta, I, D). \tag{3.16}$$

ii) The M-step maximizes

$$E_{Q_{\textbf{old}}(V)}[\log P(V, \vartheta|I, D)] \tag{3.17}$$

with respect to ϑ where $Q_{\textbf{old}}(V)$ is found from the previous VE-step.

We provide the technical details of the VE-step and M-step in the next sections, 3.2.2 and 3.2.2.

The VE-step

The true distribution of $P(V|\vartheta, I, D)$ in (3.16) is intractable to compute. Therefore, we perform the mean field approximation of $P(V|\vartheta, I, D)$ by $Q(V)$, which can be expressed as,

$$\log P(V|\vartheta, I, D) \propto \sum_{i=1}^{M} f_i(v_i) + \sum_{i=1}^{M} \sum_{j \subset \mathcal{N}_i} f_{ij}(v_i, v_j) + \sum_{i=1}^{M} \sum_{j_c=1}^{M_c} g_{ij}(v_i, v_{j_c}). \tag{3.18}$$

In this equation, $f(v_i)$ is the sum of the logarithms of the image likelihood term in (3.9) and the reconstruction error term in (3.13). $f(v_i, v_j)$ is determined by the logarithm of the compatible probability from the Potts model in (3.7). The pairwise $g(v_i, v_{j_c})$ is determined by the logarithm of the geodesic potential in (3.12), such that

$$g(v_i, v_{j_c}) = \log P_{geo}(D|v_i, v_{j_c}). \tag{3.19}$$

As in [130], the belief $q_i(v_i) = P(v_i|\vartheta, I, D)$ is iteratively updated until convergence:

$$q_{i_{step+1}}(v_i) = \frac{1}{Z_{i_{step}}(v_i)} \exp\{\sum_{j_c=1}^{M_c} \sum_{v_{j_c}} q_{step}^{j_c}(v_{j_c}) g_{ij}(v_i, v_{j_c}) + \sum_{j \subset \mathcal{N}_i} \sum_{v_j} q_{j_{step}}(v_j) f_{ij}(v_i, v_j) + f_i(v_i)\} \tag{3.20}$$

where $q_{step}^{j_c}(v_{j_c}) = E[q_{j_{step}}(v_{j_c})]$ is an average belief of all pixels $j \subset$ the cell j_c and $Z_{step}(v_i) = \sum_{v_i} q_{i_{step}}(v_i)$ is a normalization factor. To reduce the amount of computation required, set $q_{step}^{j_c}(v_{j_c} = \varepsilon) = 1$ for the maximum probability of the cell j_c pertaining to the ellipsoid ε and $q_{step}^{j_c}(v_{j_c}) = 0$ for $v_{j_c} \neq \varepsilon$. $f_i(v_i)$ is used to initialize the value of $q_i(v_i)$ in the first iteration, where

$$q_{i_0}(v_i) = \frac{1}{Z_{i_0}(v_i)} \exp\{f_i(v_i)\}. \tag{3.21}$$

In Fig. 3.9, we show the results of running the VE-step on two examples in which the label of each point is selected by the label with the maximum belief. At the first iteration, using only the image likelihood and the distance provides incorrect labeling results because some pixels belonging to an arm are near to the torso or the head. After the VE-step converges (three or four iterations), we obtain a correct labeling assignment.

The M-step

Once the distribution of the random variable v_i has been obtained, the kinematic parameter ϑ becomes the solution of the following optimization problem:

$$\text{argmax}_\vartheta E_{Q(V)}[\log P(D|\vartheta, V)]. \tag{3.22}$$

Here, the components independent of ϑ in (3.6) are eliminated. By taking the logarithm of $P(D|\vartheta, V)$, (3.22) can be rewritten as

$$-\text{argmax}_\vartheta \sum_{\varepsilon=1}^{N_\varepsilon} \sum_{i=1}^{M} q_i(v_i = \varepsilon) d^2(X_i, \vartheta, v_i = \varepsilon) \tag{3.23}$$

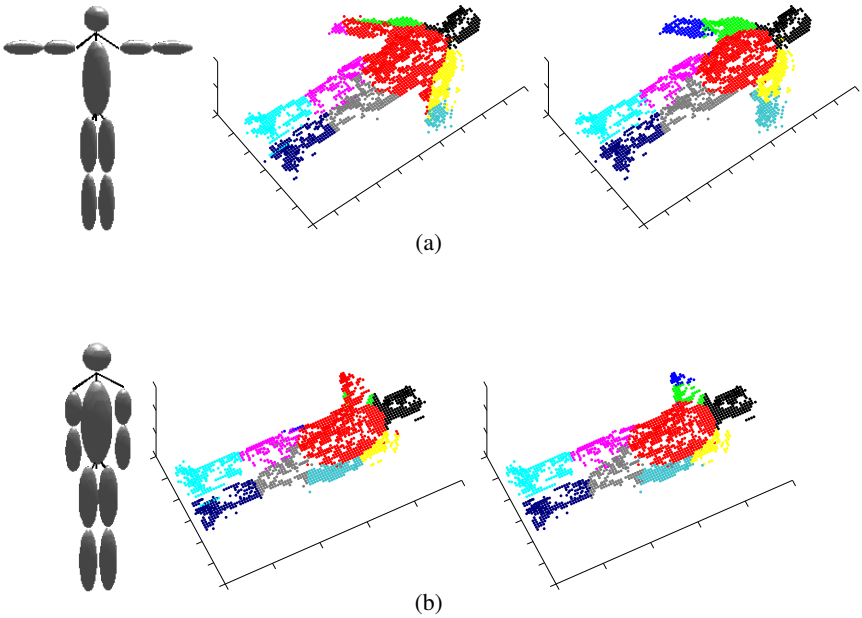


Figure 3.9: The results of running the VE-step on two examples (a) and (b). Corresponding from left to right: the initial human models, the label assignments found by the first iteration of the VE-step, and the last iteration.

where N_ε is the number of ellipsoid, $d(X_i, \vartheta, v_i = \varepsilon) = \|X_i - Z_i(\vartheta)^\varepsilon\|^2$, and $Z_i(\vartheta)^\varepsilon$ is the nearest point of X_i lying on the surface of the ellipsoid ε . We formulate (3.23) in an alternative way as

$$\text{argmin}_\vartheta \sum_{\varepsilon=1}^{N_\varepsilon} \sum_{i=1}^{M} q_i(v_i = \varepsilon) \|X_i - Z_i(\vartheta)^\varepsilon\|^2. \tag{3.24}$$

For simplification of the M-step, set $q_i(v_i = \varepsilon) = 1$ for the maximum probability that the point i pertains to the ellipsoid ε and $q_i(v_i) = 0$ for $v_i \neq \varepsilon$. The least square problem with a nonlinear function like (3.24) can be efficiently solved by the Levenberg-Marquardt method. This estimator requires the computation of the Jacobian matrix $\mathbf{J}$ of $Z_i(\vartheta)^\varepsilon$ with respect to ϑ [84, 91, 125] that is explained in Appendix E.

3.3 Chapter Summary

In this chapter, we have presented our marker-less system to recover human body poses in 3-D from depth images acquired by a single stereo camera. We have described our methodology including how to estimate the 3-D data of a depth image, how to create a human body model, and how to register the human body model to the 3-D data. We estimated the pixel displacements of stereo image pairs to reconstruct 3-D information. We modeled the human body with a set of ellipsoids connected by kinematic chains and parameterized with rotational angles at each body joint. To solve our registration problem minimizing the difference between the human model and the information in a depth image to recover a human pose, we derive an algorithm based on VEM with two-step iterations: assigning the 3-D data to different body parts and refining the kinematic parameters to fit the 3-D model to the data. The algorithm is iterated until it converges on the correct pose. The experimental results validating our proposed method are correspondingly presented in Section 5.1 and Section 5.2. Subsequently, the pose recognition is applied to estimate human poses from a sequence of depth images and the joint angles of a sequence of estimated poses are utilized in Chapter 4 to recognize different activities.

Chapter 4

Human Activity Recognition Using Body Joint Angles

A general method for video-based HAR starts with extracting key features from images and comparing them against the features of various activities. Thus, activity feature extraction techniques play important roles in this regard. In this chapter, we present how various human activities can be recognized with the new features of body joint angles derived from the body poses recovered from stereo data. The features of body joints angles are utilized over the conventional binary body silhouettes and HMMs are used to recognize various human activities. The chapter is organized as follows. In Section 4.1, we compare the process involved the binary silhouette- and joint angle-based HAR. Section 4.2 and Section 4.3 describe the characteristics of binary silhouette features and of joint angle features, respectively. An HMM with its roles in training and classifying sequential features is introduced in Section 4.4. Finally, we summarize the chapter contents in Section 4.5.

4.1 Binary Silhouette- and Joint Angle-based HAR

In general, 2-D binary silhouettes of human body shapes are the most common representations of human activity that have been applied for parametric-based HAR [24, 93, 94, 132, 133, 146] (see Section 2.3.2). The top flow of Fig. 4.1 shows the typical processing components of the binary silhouette-based HAR. Once the binary silhouettes are obtained from RGB images, some prominent features, obtained through the feature extraction process such as PCA or ICA, are then applied to a recognition technique of HMMs to train and recognize various human activities.

Recently, more advanced HAR techniques have been introduced in terms of new features and more powerful feature extraction techniques. Although binary silhouettes are commonly employed to represent a wide variety of body configurations, they also produce ambiguities by representing the same silhouette for different poses from different activities, especially for those activities that are performed toward the video camera. Thus, the binary silhouettes do not seem to be a good choice to represent human body poses in different activities.

As the human body consists of limbs connected with joints, if one is able to obtain their 3-D joint angle information, one can form much stronger features than conventional silhouette features that will lead to significantly improved HAR. From the time-sequential activity video frames, the joint angles are first estimated by co-registering a 3-D human body model to the stereo information and then mapped into codewords to generate a sequence of discrete symbols for an HMM of each activity. With these symbols, each activity HMM is trained and used for activity recognition. The bottom of Fig. 4.1 shows the basic processes regarding 3-D body joint angle-based HAR. It indicates that after obtaining the depth images, joint angles are estimated via co-registration and represented as features to feed into the HMMs to train and recognize different human activities.

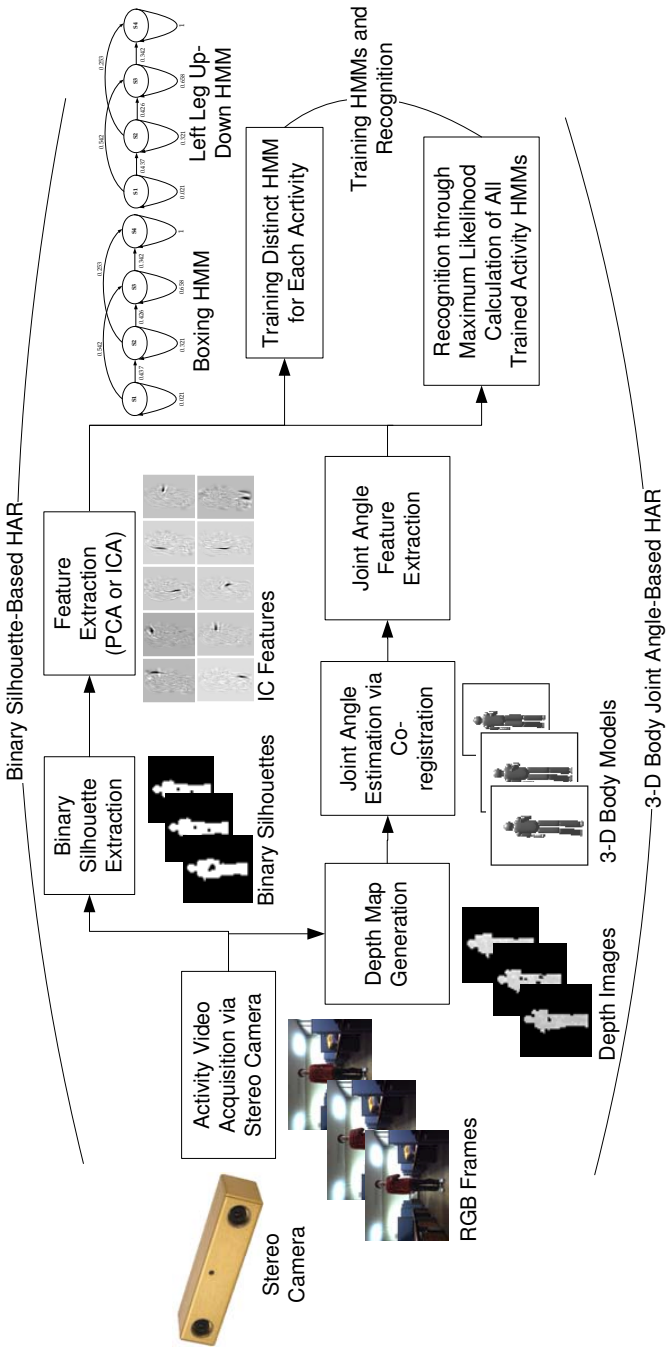


Figure 4.1: Processes involved in the binary silhouette and 3-D body joint angle-based HAR.

4.2 Binary Silhouette Features in Human Activities

This section describes the method of using PCA and ICA to extract classifying features from binary body silhouettes. Assume that each image containing a binary silhouette has a size of $h \times w$. Correspondingly, an image at the time index k is represented by a vector X_k of size $1 \times D$, where $D = hw$. There are a total of m extracted independent components (ICs) or principal components (PCs) with a size of $1 \times D$, $e_1, e_2, ..., e_m$ extracted by ICA or PCA from the training dataset. We can calculate a PCA or ICA projection F_k of X_k when projecting X_k on m extracted components $\mathbf{E} = (e_1, e_2, ..., e_m)^T$ using

$$F_k = X_k \mathbf{E}^+, \tag{4.1}$$

where $\mathbf{E}^+$ is the pseudoinverse of the extracted components. The representation F_k has a number dimension much smaller than that of X_k, consists more compact information, and is therefore used as the replaced features of binary silhouettes. A set of features corresponding to a sequence of frame $1, 2, ..., T$ is expressed by $\{F_1, F_2, ..., F_T\}$ where F_k is computed from equation (4.1).

4.2.1 Principle component analysis of body silhouettes

Given the set of training images $\mathbf{X} = (X_1, X_2, ..., X_n)^T$, the objective of PCA [99] is to find a set of bases w_i (or the weight vectors) to preserve as much information as possible of $\mathbf{X}$ when we project $\mathbf{X}$ onto the space spanned by the bases w_i. Here, the data are supposed to have a zero mean (i.e, the mean has been subtracted from the data set). The projection of X_k onto the basis w_i is given by,

$$\hat{X_k} = w_i^T \, X_k w_i. \tag{4.2}$$

As the results of conventional work [41, 95], the value of the weight vector w_i minimizing the square of the differences between X_k and $\hat{X_k}$, $\|X_k - \hat{X_k}\|^2$, is obtained by

$$\mathbf{C}_x w_i = \lambda w_i \tag{4.3}$$

where $\mathbf{C}_x = E\{\mathbf{X}\mathbf{X}^T\}$ is the covariance matrix of $\mathbf{X}$. Obviously, equation (4.2) states that w_i is the eigenvector of the matrix $\mathbf{C}_x$. In PCA, just important eigenvectors corresponding to the largest eigenvalues of $\mathbf{C}_x$ are retained to form a projection space [69, 144]. A total of m largest eigenvectors $[w_1, w_2, ..., w_m]^T$ are used to formulate the bases $\mathbf{E}_{PCA}$ of PCA. Due to the uncorrelation of extracted PCs, equation (4.1) is simplified by

$$F_k = X_k \mathbf{E}_{PCA}^{+} = X_k \mathbf{E}_{PCA}^{T}. \tag{4.4}$$

We show some PC examples of binary silhouettes estimated using our training data in Fig 4.2. Here, these PCs all present the global shape of the whole human body.

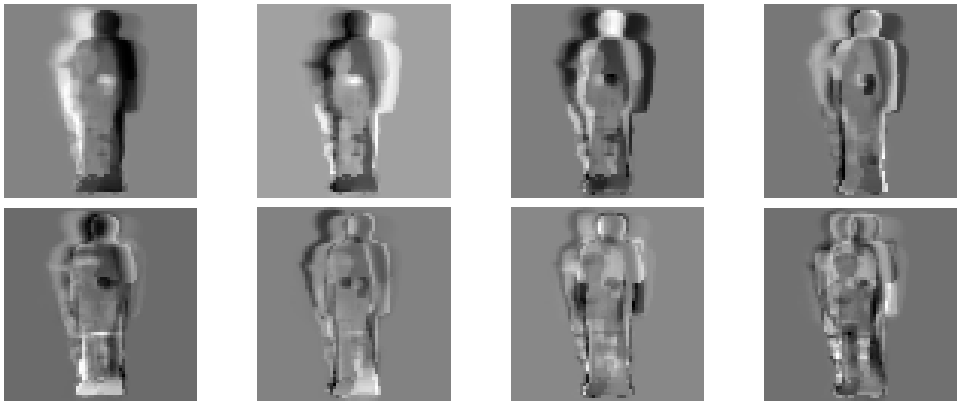


Figure 4.2: Eight PCs from all activity silhouettes.

4.2.2 Independent component analysis of body silhouettes

Assume that observed images are a linear mixture of some original sources of images. The goal of ICA is to recover the original sources from a set of training images. If the observed images of a training dataset are presented as $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ and the original sources as $\mathbf{S} = (S_1, S_2, ..., S_m)^T$, an assumption that $\mathbf{X}$ is a linear mixture of original sources is stated by

$$\mathbf{X} = \mathbf{A}\mathbf{S} \tag{4.5}$$

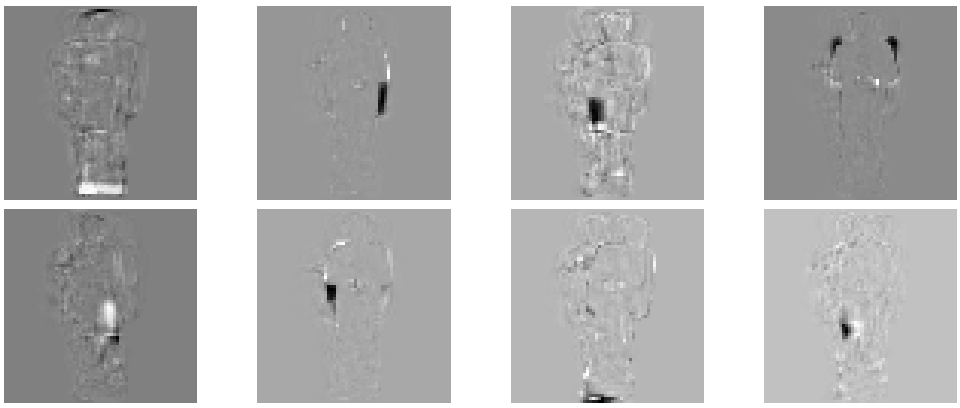


Figure 4.3: Eight ICs from all activity silhouettes.

where **A** is the mixing matrix with size $(n \times m)$. The ICA algorithm is used to compute the $(m \times n)$ demixing matrix $\mathbf{W} = [w_1, w_2, ..., w_m]^T$ to recover all original signals from the observed

$$\mathbf{Y} = \mathbf{WX} \tag{4.6}$$

where $\mathbf{Y} = (Y_1, Y_2, ..., Y_m)^T$ and each separated output image is $Y_i = w_i^T \mathbf{X}$. Usually, before mixed to form observations, the original signals tend to be mutually independent together. This is the underlying fundamental to formulate the basics of the ICA algorithm: the ICA algorithm aims at finding the demixing matrix **W** to make Y_i independent as much as possible and thus converge toward one of ICs. More essentials of ICA are provided in [65].

Usually, the ICA algorithm is performed on the m extracted PCs of the training data. Therefore, the corresponding ICs are found by

$$\mathbf{E}_{ICA} = \mathbf{WE}_{PCA} \tag{4.7}$$

Some examples of extracted ICs depicted in Fig. 4.3 reveal the fact that ICs focus more on local body components, commonly activated in human movements. Thus IC features seem better than PC features in terms of distinguishing different human activities.

For ICA, an equation to compute representation F_k of projecting X_k on the space spanned by the extracted ICs is obtained from (4.1) with some simplifications

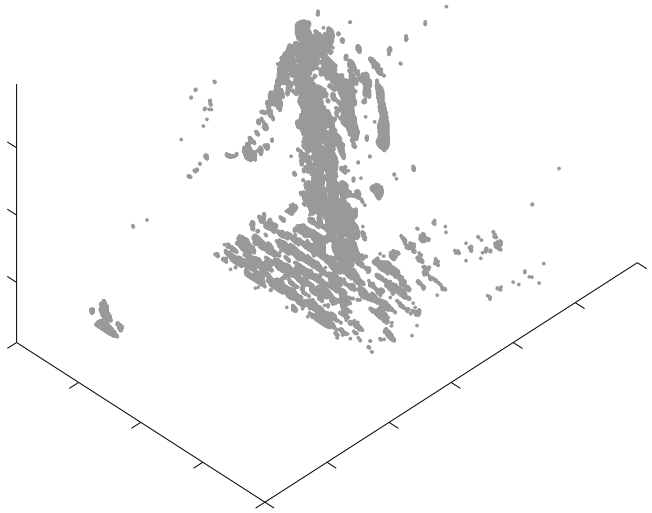
$$\begin{aligned} F_k &= X_k \mathbf{E}_{ICA}^{+} = X_k (\mathbf{E}_{ICA}^T \mathbf{E}_{ICA})^{-1} \mathbf{E}_{ICA}^T \\ &= X_k \mathbf{E}_{PCA}^T \mathbf{W}^{-1}. \end{aligned} \tag{4.8}$$

4.3 3-D Joint Angle Features in Human Activities

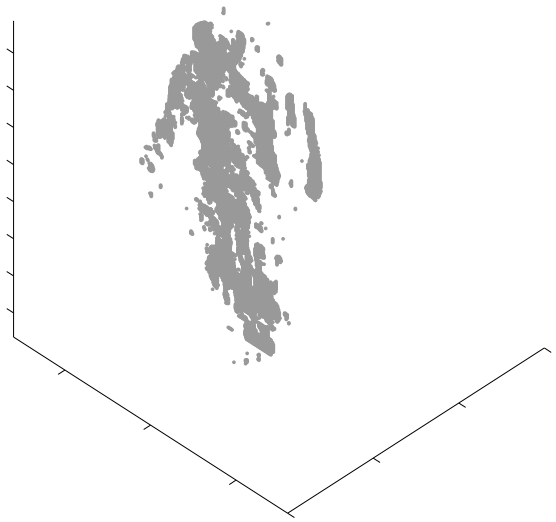
4.3.1 Location tracking of a moving subject

In Chapter 3, we developed a method to estimate human body poses from stereo data of a subject performing activities in a fixed location. In this work, concerning with recovering human body poses of a subject moving in the horizontal (e.g., walking) and vertical direction (e.g., sitting), we have added a tracking step to locate the subject's position. The subject's location is used to remove the artifacts which are a part of 3-D data remaining far from the subject as depicted in Fig. 4.4(b). Furthermore, face detection is utilized to detect the head and torso areas as depicted in Fig. 4.5 which are used in the labeling step of the co-registration algorithm. Finally, the six parameters of the global transformation from the global coordinate system to the local coordinate system at the body hip are computed with the subject's location obtained by the tracking step, giving higher precision.

Let a pair of parameters $[s_t^H, s_t^B]$ present the location of a human subject where s_t^H and s_t^B are the two 3-D vectors locating the center of the head and the body at the time index t. From the information of RGB images and 3-D data, we can obtain the approximate values of $[s_t^H, s_t^B]$ by $[r_t^H, r_t^B]$: we detect the head region from RGB images and 3-D data by the face detection algorithm using the Haar features [135] to compute r_t^H; We track the body region from RGB images and 3-D data using the Mean shift algorithm [32, 33] to get a value of r_t^B. Let $\dot{s}_t^H$ and $\dot{s}_t^B$ be the velocity of the head and body at the time index t. A set of equations established to track the changes from $[s_{t-1}^H, s_{t-1}^B]$ to $[s_t^H, s_t^B]$ and the relationship between the real human location

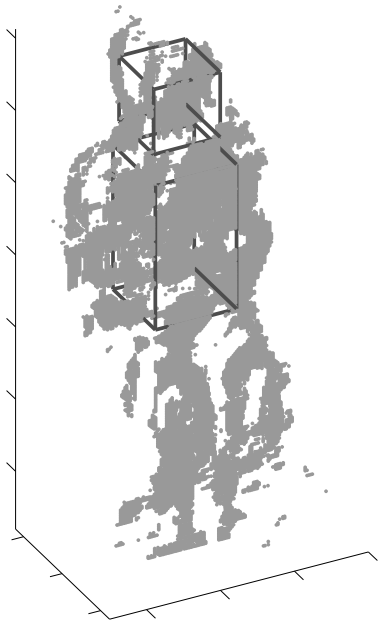


(a)



(b)

Figure 4.4: A sample of (a) 3-D data of a moving person, (b) a noise removal of 3-D data of a moving subject.



(a)

Figure 4.5: Detecting head and torso of a sitting person.

$[s_t^H, s_t^B]$ and the raw estimation $[r_t^H, r_t^B]$ is given by

$$s_t^H = s_{t-1}^H + \dot{s}_t^H \tau + \upsilon_1 \tag{4.9}$$

$$s_t^B = s_{t-1}^B + \dot{s}_t^B \tau + \upsilon_2 \tag{4.10}$$

$$\dot{s}_t^H = \dot{s}_{t-1}^H + \upsilon_3 \tag{4.11}$$

$$\dot{s}_t^B = \dot{s}_{t-1}^B + \upsilon_4 \tag{4.12}$$

$$r_t^H = s_t^H + \zeta_1 \tag{4.13}$$

$$r_t^B = s_t^B + \zeta_2 \tag{4.14}$$

$$d = \|s_t^H - s_t^N\| + \zeta_3 \tag{4.15}$$

where $\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \zeta_1, \zeta_2$ and ζ_3 are random variables drawn from a Gaussian distribution, τ the time interval between two frames, and d the constant distance between the center of the head and the center of the body. We update the current subjects' position $[s_t^H, s_t^B]$ from the previous estimation $[s_{t-1}^H, s_{t-1}^B]$ and from the observation $[r_t^H, r_t^B]$ by Extended Kalman Filter [108]. The face and torso regions are then estimated from $[s_t^H, s_t^B]$ by the method presented in [128].

4.3.2 Human pose estimation and joint-angle feature extraction

A step-by-step processing to extract body joint-angle features of a stereo sequence is depicted in Fig. 4.6. Given a sequence of stereo image-pairs $(S_1, S_2, ..., S_T)$ where T is the length of a video activity, the depth data of each stereo frame k are estimated by a method illustrated in Section 3.1.1 of Chapter 3. Additionally, to recover the poses of a moving subject, a tracking algorithm in Section 4.3.1 is applied to locate the changing position of the subject. The step to recover 3-D human poses from depths is presented in Section 3.2 of Chapter 3: we define our 3-D human model with a set of connected ellipsoids which are parameterized by kinematic angles; The angular kinematic angles are adjusted to fit the 3-D model to the observation; Consequently, we can reconstruct the human poses reflected in stereo images.

As denoted in Section 3.1.2 of Chapter 3, 24 kinematic parameters $(\vartheta_1, \vartheta_2, ..., \vartheta_{24})$ of joint

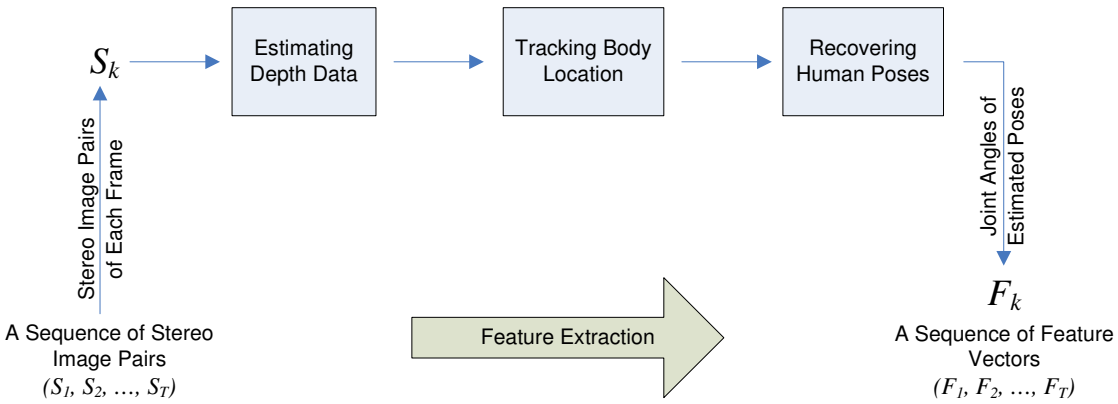


Figure 4.6: Basic steps of estimating body joint angles of a stereo sequence.

angles are used to model a specific pose. Once we successfully obtain the 3-D human pose for each video frame, we can utilize its joint angles to represent various human activities effectively. The estimated joint angles from a video frame of a particular activity form a feature vector $F_k = (\vartheta_1, \vartheta_2, ..., \vartheta_{24})$. Thus, each activity video clip is represented in a sequence of joint angle feature vectors as $(F_1, F_2, ..., F_T)$. Therefore, the 3-D joint angle features from video can really contribute in distinguishing an activity from another, especially those activities that are not discernible with the conventional binary or depth silhouette-based approaches.

4.4 Training and Recognition via HMM

HMM has been applied extensively to solve a large number of spatiotemporal pattern recognition problems including human activity recognition because of its capability of handling sequential information in space and time with its probabilistic learning capability for recognition [74, 93, 94, 132, 133]. Basically, an HMM is a stochastic process where an underlying process is usually unobservable but it can be observed through another set of stochastic processes that produces observation symbols. The graphical structure of an HMM and how to perform probabilistic

inference on an HMM are described in Appendix B. To learn a video-based human activity in an HMM, the symbol sequences obtained from the training image sequences of distinct activities are used to optimize the corresponding HMM. Finally, the trained HMMs are used to calculate the maximum likelihood for recognition.

Technically, HMM is a collection of finite states connected by transitions. Every state is characterized by transition and symbol observation probabilities. A generic HMM is expressed as $H = \{S, \pi, A, B\}$ where S denotes possible states, π the initial probability of the states, A the transition probability matrix between the hidden states and B the observation probability from every state. If the number of activities is N then there will be a dictionary $(H_1, H_2, ..., H_N)$ of N trained models. To estimate HMM parameters, one could use the Baum-Welch algorithm [74].

We choose a four-state and left-to-right HMM in this study to model sequential events of each human activity. To recognize each test activity, the obtained observation symbol sequence $O = \{O_1, O_2, ..., O_T\}$ through the vector quantization process is used to determine the proper activity HMM from all the trained activity HMMs by means of the highest likelihood as

$$decision = \text{argmax}_{i=1,2,...,M}\{P(O|H_i)\} \tag{4.16}$$

where H_i indicates i^{th} HMM and M number of activities. More details on regarding training and testing of HMMs for human activity recognition are available in our previous work [132, 133].

4.5 Chapter Summary

We describe how the body poses estimated from depth images and their derived parameters (i.e., joint angles) can be used for HAR. We introduce the conventional methods using binary silhouettes with PCA or ICA for feature extraction. For comparison, we presented experiments performed on hundreds of video sequences in Section 5.3 to validate whether the presented techniques outperform the conventional techniques of HAR using binary silhouette features.

Chapter 5

Experimental Results

We evaluate our techniques proposed in Chapter 3 to recover human poses from stereo images with simulated and real data in Section 5.1 and 5.2. Then, we present our constructed database consisting of hundreds of video sequences and their recovered pose sequences of different activities in Section 5.3. We compare the recognition rates our joint angle-based HAR proposed in Chapter 4 with binary silhouette-based HAR on an activity database in Section 5.4.

5.1 Experimental Results of Estimating Human Poses from Simulated Stereo Data

In generating the simulated data, we manually defined some joint angle trajectories as depicted by the dashed lines in Fig. 5.2. Only the rotational angles corresponding to the elbow and shoulder joints were tested in our experiments; the values of other rotational angles were fixed. From the predefined angle trajectories, we created a sequence of human poses and their disparity images up to 110 frames. Some samples of the disparity images are shown in the first and third rows in Fig. 5.1.

We applied our algorithm to recover the human poses from the synthetic disparity images.

Due to the nature of simulated data, the cues of RGB images were not available, so we eliminated them from computation. Some samples of the recovered human poses are depicted in the second and fourth rows in Fig. 5.1. To validate our algorithm, we plotted the estimated angle trajectories as solid lines to compare against the synthetic angle trajectories plotted as the dashed line in Fig. 5.2. The results show the good estimation of the kinematic parameters achieved by our method.

5.2 Experimental Results of Estimating Human Poses from Real Stereo Data

Experiments were implemented with stereo data acquired by the stereo camera. In Section 3.1.1, we described the use of the GCS algorithm to extract the disparity image and compute the 3-D data for each frame. The subjects were asked to perform some distinguishable activities about 2-4 meters from the camera, producing several video sequences. The reconstructed body poses were validated by visually checking the trajectories of certain joint angles.

In the first experiment, we assessed the movements of elbows in both horizontal and vertical directions, as shown in Fig. 5.3. In each figure, the sequence of activities is illustrated in a video stream from top to bottom in a column. Observing the real pictures, the angle changes between the upper arm and lower arm were approximately 90^0. In Fig. 5.4, the recovered angle of the second joint precisely reflects the arm motion in the real data. The joint angles may receive positive or negative values, depending on the way that two joint angles at the elbow are combined to drive the arm movements.

In the next test dataset, as shown in Fig. 5.5(a), the activity of the person in the video was related to the movements of the knee joint. The right leg was lifted until it made a 90^0 angle between the upper leg and lower leg, then this was followed by the same motion of the left leg. The kinematic motion parameters were estimated and are depicted in Fig. 5.7(a). One may notice that the switching between the two legs happens from frame 70 to frame 80.

In order to track the changes of two joint angles at the same time, we considered the sequence of activities in Fig. 5.5(b). We assumed that the whole arm laid along the x-axis and that the two joint angles of the shoulder were related to the rotation of the arm around the z-axis and x-axis, respectively. One can observe both trajectories of the two measured joint angles from Fig. 5.7(b), with the upper curves reflecting the rotational angles around the z-axis and the lower curves reflecting the rotational angles around the x-axis. To explain the meaning of the plot, we visualize the overall progress in Fig. 5.6. First, the whole arms were rotated around the z-axis from 180^0 to 360^0 $(+\pi)$, corresponding to the vertical movement within the frames 1-45. At the second stage, the second joint angles changed their values from 180^0 to 270^0 $(+\pi/2)$, while the arms retained their positions from frames 45 to 60. Finally, to be horizontally extended to the left or right side, the two arms were continuously rotated around the z-axis (the first joint angles) from 360^0 to 270^0 $(-\pi/2)$ or 450^0 $(+\pi/2)$, corresponding to frames 60 to the end.

To quantitatively evaluate the reconstruction errors of these experiments, we needed to generate ground-truth using the given data. Applying the same method presented in [56, 76], the locations of some distinct points (e.g., hands, elbows, or shoulders) were hand-labeled in the RGB images. We used the 3-D information from these points to calculate the necessary ground-truth angles between two limbs. The angles reconstructed by the kinematic parameters were compared against the ground-truth by the average error ϵ_ϑ

$$\epsilon_\vartheta = \frac{\sum_{t=1}^{n} |\vartheta_t^{est} - \vartheta_t^{grd}|}{n} \tag{5.1}$$

where n is the number of frames, t is the frame index, ϑ_t^{grd} is the ground-truth, and ϑ_t^{est} is the estimated angle. In particular, the shoulder movements were related to two kinematic parameters, and therefore the correct arm directions were validated by measuring the angles between the arms and the x-axis or z-axis. The coordinate system (x, y, z) in this case had the x-axis and z-axis aligned with the vertical and horizontal directions of the image plane, respectively. The average errors of all four experiments are given in Table 5.1.

Fig. 5.8 shows the results of testing our algorithm on some free movements. The subjects

performed complicated activities with all of their arms and legs. Here, we depict only three images out of the sequence and their estimated poses in the second and third rows with two alternative view angles. The 3-D locations of the body parts and the correct human poses were successfully identified. In these experiments, it is more convenient to evaluate the estimated whole body pose, rather than the local changes of individual limbs. The average distance between each 3-D point and the nearest ellipsoid of the reconstructed model can be considered the overall error measurement of the reconstructed pose in each frame. The average distance D_t of the frame t is computed by

$$D_t = \frac{\sum_{i=1}^{M} d_t(i)}{M} \tag{5.2}$$

where $d_t(i)$ is the Euclidean distance from the point i to the nearest ellipsoid and M is the number of points. The means and standard derivations of D_t in the two last sequences are provided in Table 5.2.

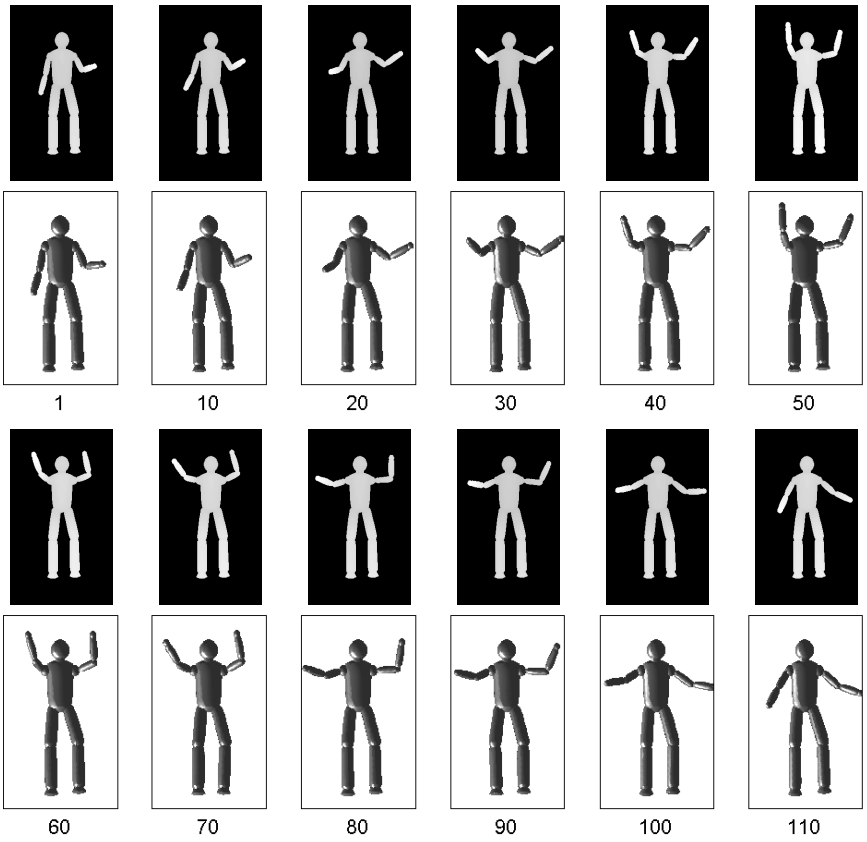


Figure 5.1: The results of recovering human poses (the second and fourth rows) from the synthetic disparity images (the first and third rows). The number below each picture indicates the frame index number.

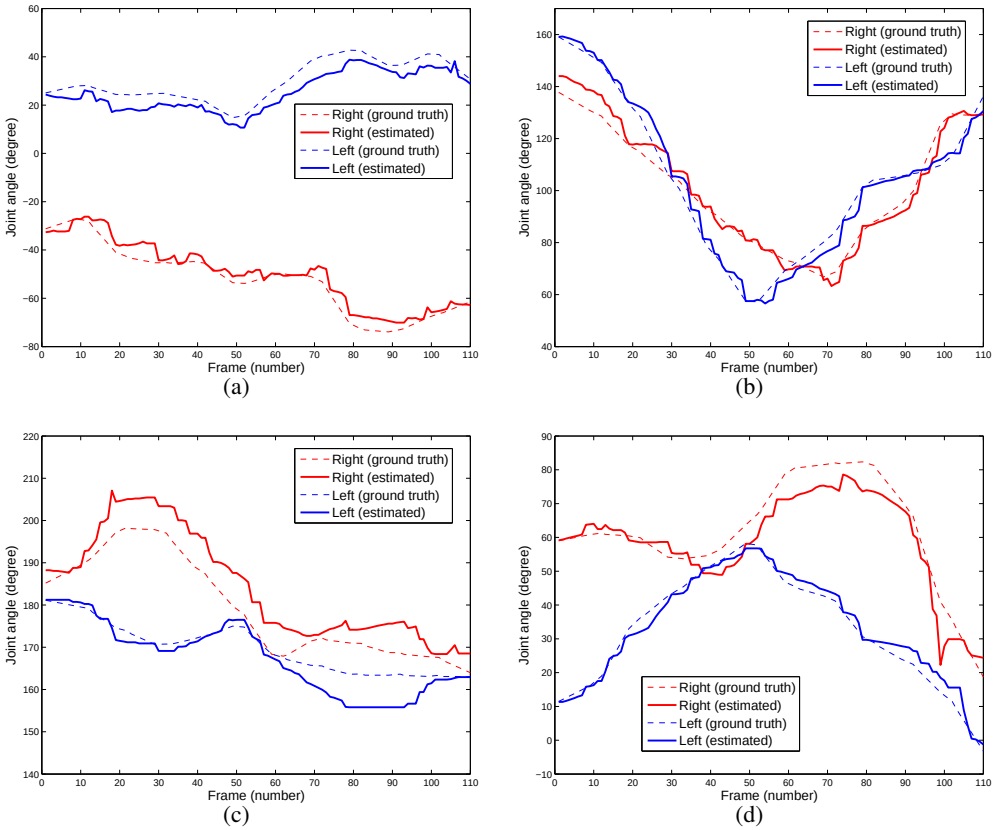
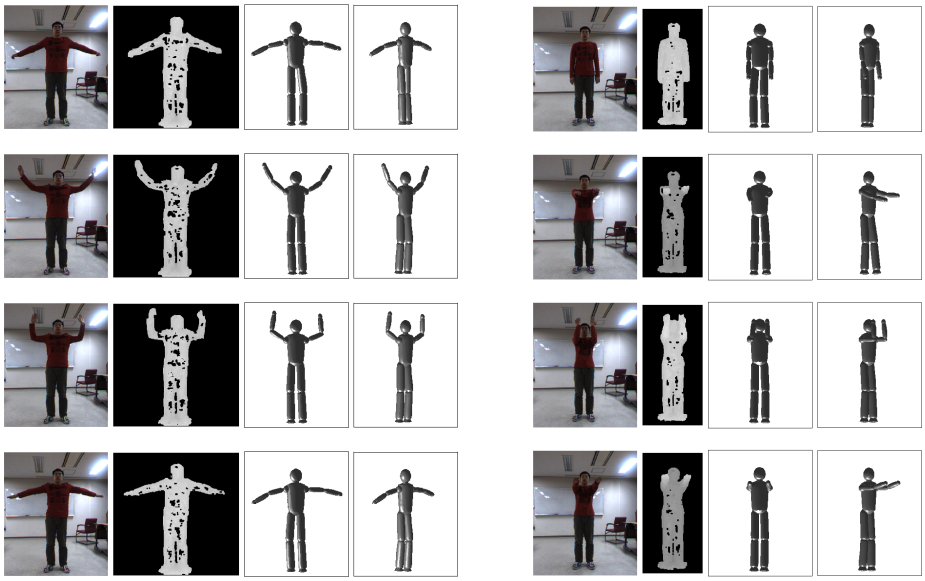


Figure 5.2: A comparison between the estimated and the ground-truth joint angles in the simulated experiments (synthetic data). (a) and (b) show two joint angles of the shoulders. (c) and (d) show two joint angles of the elbows.



(a) (b)

Figure 5.3: Real experiments with elbow motion in two different directions. (a) Horizontal movements. (b) Vertical movements. From left to right: the RGB images, disparity images, and reconstructed human models (front view and $+45^0$ view).

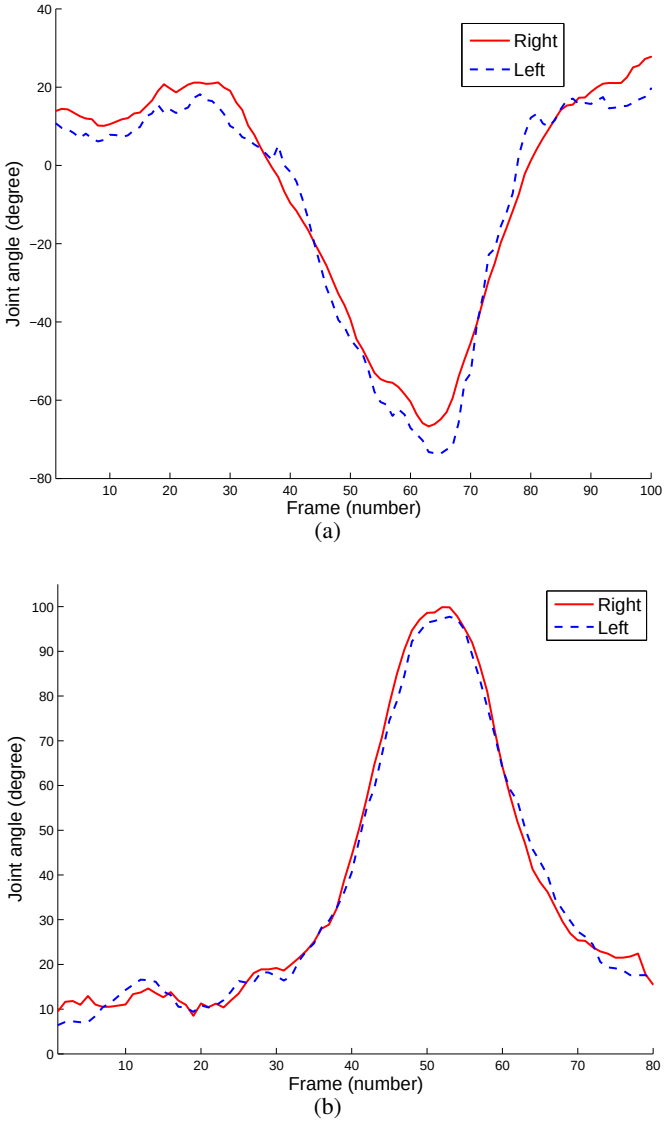
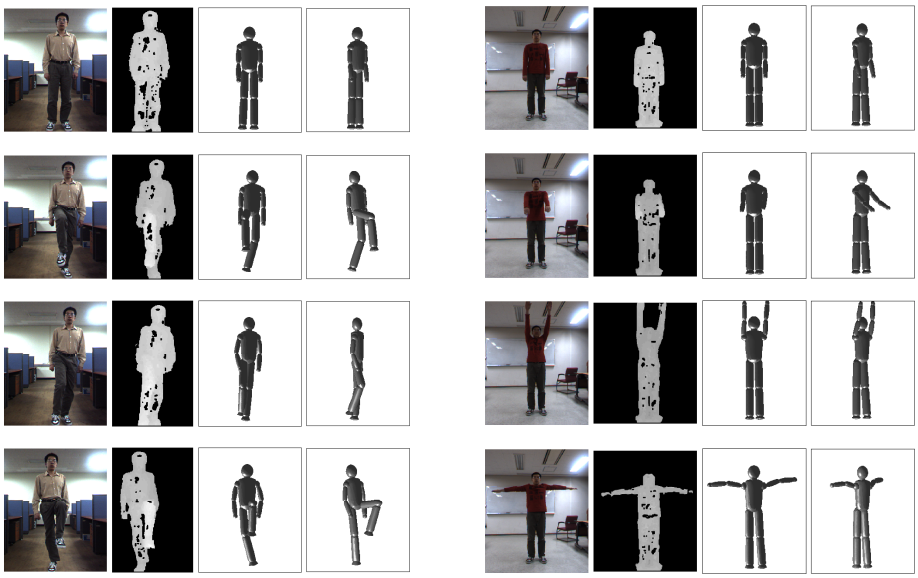


Figure 5.4: The estimation of the second joint-angle trajectories for the left and right elbows corresponding to: (a) horizontal elbow movement and (b) vertical elbow movement.



(a) (b)

Figure 5.5: Real experiments with other motions: (a) Knee movements. (b) Shoulder movements. From left to right: the RGB images, disparity images, and reconstructed human models (front view and $+45^0$ view).

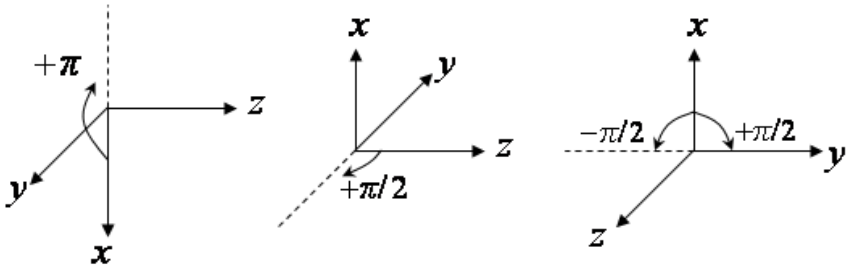


Figure 5.6: The changes in two joint-angles during the movements of the shoulders (experiment depicted in Fig. 5.5(b)).

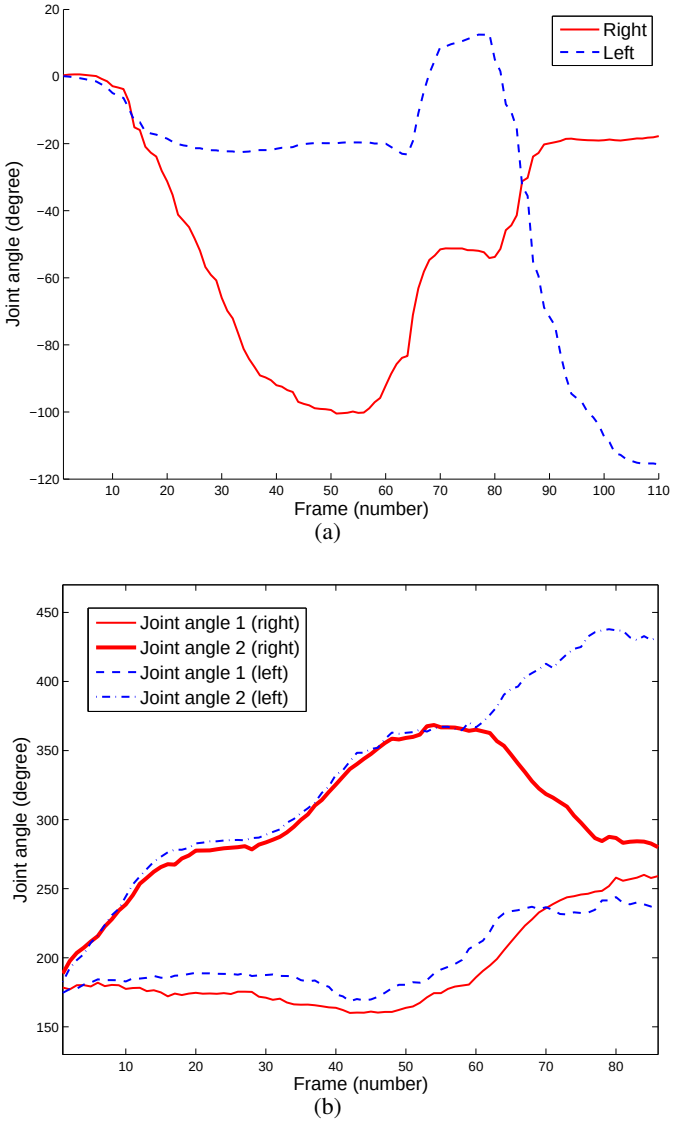


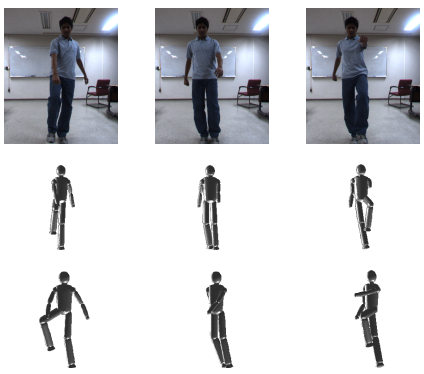
Figure 5.7: The estimation of the joint-angle trajectories for the left and right sides of: (a) knee movements and (b) shoulder movements.

Table 5.1: The average reconstruction error (^{0}) of the joint angles of the first four experiments. Note that these experiments only consider the local movements of some body limbs.

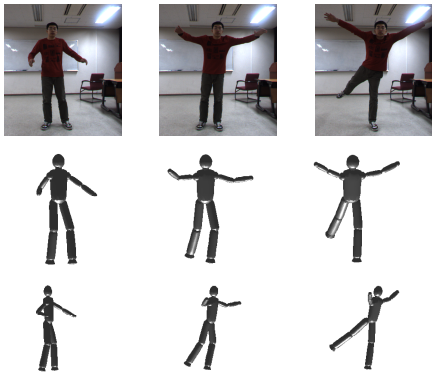
Experiment	Evaluated angle	Average reconstruction error	
Elbow movement (horizontal direction)	Upper arm & lower arm	Left	8.21
		Right	7.58
Elbow movement (vertical direction)	Upper arm & lower arm	Left	6.79
		Right	7.64
Knee movement	Upper leg & lower leg	Left	8.03
		Right	13.81
Shoulder movement	Whole arm & x-axis	Left	5.66
		Right	5.72
	Whole arm & z-axis	Left	9.08
		Right	9.97

Table 5.2: The mean and standard derivation of the average distance (the average Euclidean distance between a set of 3-D points of the observed data and the ellipsoids of the reconstructed model) of the last two sequences.

Sequences	Walking	Arbitrary activity
Mean (m)	0.062	0.037
Std. Dev. (m)	0.003	0.002



(a) Reconstructed human body poses with the front view and -45^0 view.



(b) Reconstructed human body poses with the front view and $+45^0$ view.

Figure 5.8: The qualitative evaluation of the reconstructed human body poses from: (a) walking sequences and (b) arbitrary activity sequences.

5.3 Human Activity Database

We built a database of eight different activities (namely, left hand up-down, right hand up-down, both hands up-down, boxing, left leg up-down, right leg up-down, walking, and sitting) to be trained and recognized via our 3-D joint angle and HMM-based approach. A total of 15 and 40 image sequences of each activity were prepared to be used for training and recognition respectively. Some samples of pose sequences estimated from different activities are depicted in Fig. 5.9.

5.4 Experimental Results of Recognizing Various Human Activities with Joint Angle-based HAR and Binary Silhouette-based HAR

We started our experiments with the traditional binary silhouette-based HAR. Table 5.3 and 5.4 show the experimental results of HMM-based HAR utilizing the PC and IC features of binary silhouettes, respectively. Table 5.5 shows the experimental results of HMM-based HAR using joint angle features of 3-D body model. We consider 150 features in the feature space of both PCA and ICA-based approaches. Binary silhouettes were not appropriate to recognize the activities used in our experiments, yielding a much lower mean recognition rate of 58.12% for PCA and 64.06% for ICA, as ICA is superior to PCA by extracting the local binary silhouette features [132]. On the contrary, utilizing the 3-D body joint angle features, we obtained a mean recognition rate of 92.81%, which is far better than that of the binary silhouette-based HAR. The experimental results show that the 3-D joint angle features are remarkably superior to the conventionally used silhouette features. The body joint angle features seem to be much more sensitive toward complex activities that are not discernable with the body silhouettes.

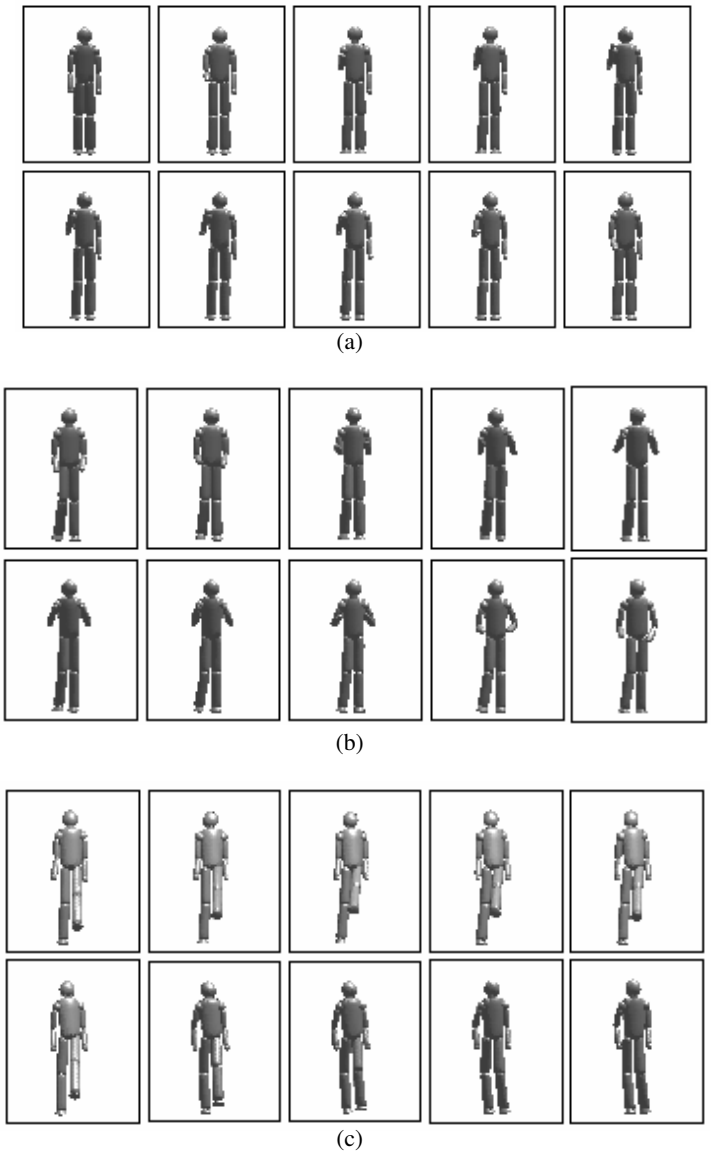


Figure 5.9: Samples of pose sequences estimated from (a) right hand up-down (b) both hands up-down, and (c) left leg up-down activities.

Table 5.3: Experimental results of PCA-based HAR using binary silhouette features.

Activity	Recognition Rate(%)	Mean	Standard Deviation
Left hand up-down	47.50	58.12	19.03
Right hand up-down	55		
Both hands up-down	60		
Boxing	20		
Left leg up-down	60		
Right leg up-down	67.50		
Walking	70		
Sitting	85		

Table 5.4: Experimental results of ICA-based HAR using binary silhouette features.

Activity	Recognition Rate(%)	Mean	Standard Deviation
Left hand up-down	47.50	64.06	18.03
Right hand up-down	60		
Both hands up-down	67.50		
Boxing	30		
Left leg up-down	72.50		
Right leg up-down	72.50		
Walking	75		
Sitting	87.50		

Table 5.5: Experimental results of HAR using 3-D joint angle features.

Activity	Recognition Rate(%)	Mean	Standard Deviation
Left hand up-down	87.50	92.81	3.65
Right hand up-down	97.50		
Both hands up-down	87.50		
Boxing	95		
Left leg up-down	92.50		
Right leg up-down	95		
Walking	92.50		
Sitting	95		

Chapter 6

Conclusion and Future Researches

6.1 Conclusion

6.1.1 Thesis summary

Developing an automatic system to extract information of people from video or images remains challenging in computer vision. So far, many studies have focused on a human pose recognition system to acquire an articulation of a human body. But understanding human is not just locating relative positions of the body limbs specified by a pose. Once the poses of a person changing overtime are known, the information about the body part motion is subsequently available to infer what people is doing. Thus combining pose and activity recognition in an engine allows us to obtain more valuable information about human states.

In this thesis work, we implement a system to recognize both human poses and activities from depth images acquired by a single stereo camera. Previously, these two tasks are typically done with a system using optical markers. Such a system is capable of producing kinematic parameters of human motion with high accuracy and speed. However, a user needs to wear specially designed optical markers when running this system. Close to our approach, a single monocular camera is utilized to capture video data. However, monocular images may not provide enough

information to recover a precise 3-D pose due to ambiguity and occlusion. A makerless system to recognize human poses and activities from a stereo camera somehow presents several advantages over previous systems and is receiving increasingly interested.

Reconstructing a 3-D human body pose from depth images recorded by the stereo camera is implemented in this work by the *parametric-based* approach. With the addition of hidden variables (i.e., body part labels), we formulate the technique in a probabilistic reasoning framework with the combination of various potentials from depths and RGB images. The joint posterior distribution of kinematic parameters and hidden variables contains various probabilistic elements including smoothness prior, image likelihood, geodesic distance constraint, and reconstruction error. Here, the *smoothness prior* presents the pair-wise probabilistic relationships of each 3-D point with its neighbors to reduce artifacts. Some body parts able to be detected in RGB images and 3-D data provide extra information about the label of 3-D points. This information is given by the *likelihood term*. If the geodesic distance be a shortest path distance in a graph using the Dijkstra's algorithm, the pairwise geodesic relationship establishes *geodesic distance constraints* of each pair of 3-D points. Two 3-D points with two corresponding labels that disregard these constraints (i.e., too close or too far) are penalized to decrease the probability. Finally, the *reconstruction error* measures the errors (Euclidean distance) between the ellipsoids of model and the 3-D data of depths.

Obviously, the pose most suitable with the observed data will correspond to the kinematic parameter that maximizes the defined posterior probability. Here, VEM algorithm is used for our optimization problem with the appearance of hidden variable and is derived by a co-registration with two main steps:

- **VE-step** estimates the posterior distribution of hidden variables. An exact expression of this distribution is intractable to compute and is thus approximated by variational inference method.

- **M-step** (Model fitting) minimizes the reconstruction error between the model and the cloud

of 3-D points that is solved by the Levenberg-Marguardt least square estimator.

The co-registration is iterated to minimize the differences between the 3-D model and the observed data. Finally, it recovers the correct human pose with the estimated joint angles. Through experiments using synthetic and real data, we demonstrated that our algorithm can reconstruct human body pose from stereo video even for complicated movements. Analyzing the performance, we detected an average error of about 6-14^0 of the estimated kinematic angles and an average distance (i.e., difference) of about 0.04-0.06m between the reconstructed body model and the given 3-D data.

At the lower levels of understanding people in video images, our pose recognition system provides the summarization of the movements of body limbs over time. To gain an understanding of people in video images, a higher sematic level of information is acquired with a HAR engine. In our HAR, we first derive the joint angles by co-registering our 3-D body model to the depth information. Then we map the joint angles into codewords, generating a sequence of discrete symbols for an HMM of each activity. With these symbols, each activity-HMM is trained and used for activity recognition. The experimental results of our system tested with different human activities from real video sequences have shown that our approach is capable of recognizing human activities with high accuracy, about 93% in the recognition rate. This is significantly better than the conventional approaches using binary silhouettes to recognize human activities could achieve.

6.1.2 Contributions

We proposed a new system to exploit information including poses and activities of a person in stereo images. The system is developed with the *parametric-based* approach, which does not require us to generate and maintain a large database of 3-D human body poses and of human activity templates.

We show how to formulate a probabilistic registration framework consisting of hidden variables of body part labels, depths, and cues from RGB images to estimate human poses. We defined

various probabilistic elements that reflect the likelihoods of hidden variables given RGB image observation, the conditional relationships among hidden variables, and the likelihoods of joint angles given depth observation. The presented framework is robust and generic; any useful information for locating the body parts can be flexibly integrated into this framework to improve the accuracy of recovering poses. Conventional method maximizing the probabilistic formulation by maximum likelihood with marginal integration over hidden variables is limited with a simple distribution. We suggest the use of an approximate method with VEM for the complicated probabilistic inference task.

We have presented a HAR work using the derived body joint angles. The proposed HAR is able to recognize various human activities with a recognition performance outperforming that of conventional human activity recognition techniques in which binary silhouettes are utilized. In overall, the entire proposed system to acquire human poses and activities from depth images is well suited to many practical applications.

6.2 Future Researches

The thesis is concluded with the discussions of open research directions expanded from the proposed system recognizing human poses and activities from data acquired by a stereo camera.

6.2.1 Future researches of human pose recognition

Our human pose capturing system using a stereo camera is potentially applicable to many areas. However, due to existing errors of recovered kinematic angles, our system might face difficulty with practical applications requiring high accurate results of estimating motion. The other difficulty of our method relates to estimating human motion from tricker movements or rapid changes of trackers' locations. In this situation, there are large variations of the human poses between two consecutive frames. A part of information used to assign the label of 3-D data might get inaccurate, causing a missing calculation of some body parts. For such reasons, we plan our future

work to improve the reliability of our presented techniques and its robustness to handle the rapid and complex changes of human poses in a video sequence. These targets could be achieved by discovering more efficient likelihoods for body part detection, applying hierarchy registration to reduce registration time, and concerning biomedical constraints as well as temporal information to better guess a pose appearance in images.

Exploiting the likelihood of body part appearances using training data

Determining the likelihood of human body parts from an RGB image is even hard for human due to the arbitrary appearances of human poses and the fusions of different body parts in images. The arisen question is how we can perform the detection of body parts using only the cues from a depth image, which provides even less information than an RGB image does. One way to make this feasible is to learn the probability distribution of the body part appearances over an image from training data. Actually, the body part likelihood of an area has not just been dependent from visual features illustrated in that area but also from others. For instance, the areas relatively left and behind other areas of a depth image get a high change to belong to a left leg rather than other parts of the human body. Using this underlying idea, a set of depth images with the ground truths of body part locations is used to learn the high-order relationships between different areas and is further applied to estimate the body part likelihoods in a single stereo image.

Hierarchy registration to reduce processing time of fitting an articulated model to stereo data

In our proposed algorithm, a large number of kinematic parameters processed in the algorithm slow down the co-registration process. To mitigate this problem, we suggested a way of hierarchy registration and computing the kinematic parameters with a small number of points. The strategy to mitigate this problem involves an observation that there always exists a subset of 3-D points sharing the same group of kinematic parameters. Thus, rather than all of kinematic parameters

are simultaneously estimated, the kinematic parameters will be sequentially found in order. The parameter that contributes to the largest number of points is first calculated. Continuously, the parameter being the second in terms of number of points related to this parameter is computed. The similar process is repeated until all of kinematic parameters are found. In this regard, the first group affecting the rigid transformation that includes six DOF to transform the whole human body from the global coordinate system to the local coordinate system at the hip should be estimated first. Then, a kinematic chain of parameters of non-rigid transformations is consequently computed to deform the human body shape in 3-D.

Biomechanical constraint and temporal information of kinematic angles

Various biomechanical constraints on kinematic angles of the human body can be established to limit nonphysical configurations of estimated poses. Besides, if taking into account the temporal information about the changes of kinematic angles from the current to the next frame [44, 134], we can improve the accuracy of our algorithm to estimate human body poses and to deal with the artifacts of stereo data. Concerned new integrations of temporal and biomechanical constraints, the probabilistic formulation of our proposed method to recover human poses becomes complicated with high-order relationships among kinematic parameters. Thus, the inference method with VEM presented in our work must be generalized by Variational Bayesian Expectation Maximization (VBEM) [81, 143, 12, 52, 48] to approximate the distributions of kinematic parameters.

6.2.2 Future researches of HAR

As the HAR method developed in this thesis work is able to obtain high recognition rates of short time activities, we plan our future work to extend the proposed method to recognize long-term human activities within complex situations.

Long-term HAR

A reasoning engine to interpret a long-term HAR is implemented using the logging data of human activities for a long duration. Here, a conventional HAR method like the one proposed in our work is applied to recognize the primitive activities (i.e, a short time activity) of a monitored person and stores the recognition results on a repository. Mining the logging information, a knowledge-and-logic based engine [45, 66, 86, 112] allows us to infer the high-level structures and semantics of human activities such as the habits and the preferences of a user. Such information can be utilized in human activity prediction, which plays a certain role in the future developments of proactive computing.

HAR in complex environment

In general, a HAR system must deal with complicated situations where a group of individuals may enter into a system. Correspondingly, a method to identify a new person and to distinguish different activities performed by separated persons should be considered. Also, due to the movements of the human body in 3-D, the overlapping and occlusions of human body parts is difficult to avoid. A good tracking algorithm to track the trajectories of human in 3-D and to reconstruct the human poses even with occlusions is necessary in this situation. Although the appearances of multiple individuals in a scene make HAR become harder, with appropriately use, the information about the human-human and human-context interactions on contrary provides valuable cues to better recognize human activities and thus should be concerned in future researches of HAR.

Publications

Journal Publications

1. **N. D. Thang**, T. Rasheed, Y.-K. Lee, S. Lee, and T.-S. Kim, Content-based facial image retrieval using constrained independent component analysis, *Information Sciences*, 2011. (accepted)

2. M. D. Uddin, **N. D. Thang.**, J. T. Kim, and T.-S. Kim, Human activity recognition using body joint angle features and hidden Markov model, *ETRI Journal*, 2011. (accepted)

3. **N. D. Thang**, T.-S. Kim, Y.-K. Lee, and S. Lee, Estimation of 3-D human body posture via co-registration of 3-D human model and sequential stereo information, *Applied Intelligence*, 2010. 10.1007/s10489-009-0209-4.

Book Chapter

1. **N. D. Thang**, M. D. Uddin, Y.-K. Lee, S. Lee, and T.-S. Kim, Recovering 3-D human body postures from depth maps and its application in human activity recognition. *Depth Map and 3D Imaging Applications*, IGI Global Publisher, 2011. (to be published)

Conference Publications

1. **N. D. Thang**, T.-S. Kim, Y.-K. Lee, and S. Lee, Fast 3-D human motion capturing from stereo data using Gaussian clusters. In: *Proceedings of the International Conference on Control Automation and Systems (ICCAS)*, Korea, 2010.

2. **N. D. Thang**, Y.-K. Lee, Y.-H. Kim, and T.-S. Kim, Makerless 3-D human motion capturing using a stereo camera. In: *Proceedings of the Second Joint ESMAC/GCMAS Meeting (JEGM)*, US, 2010.

3. J.-H. Kim, **N. D. Thang**, and T.-S. Kim, 3-D human motion tracking and gesture recognition using a data glove. In: *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE)*, Korea, 2010.

4. **N. D. Thang**, S. Lee, and Y.-K. Lee, Fast constrained independent component analysis for blind speech separation with multiple references. In: *Proceedings of the International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, Korea, 2010.

5. M. D. Uddin, **N. D. Thang**, and T.-S. Kim, Human activity via 3-D joint angle features and hidden Markov model. In: *Proceedings of the International Conference on Image Processing (ICIP)*, China, 2010.

6. L. T. Vinh, **N. D.Thang**, and Y.-K. Lee, An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information. In: *Proceedings of the Annual International Symposium on Applications and the Internet (SAINT)*, Korea, 2010.

7. J.-H. Kim, **N. D. Thang**, H. S. Suh, T. Rasheed, and T.-S. Kim, Forearm motion tracking with estimating joint angles from inertial sensor signals. In: *Proceedings of the International Conference on Biomedical and Informatics (BMEI)*, China, 2009.

8. **N. D. Thang**, P. Truc, Y.-K. Lee, S. Lee, and T.-S. Kim, 3D-human pose estimation from 2-D depth images. In: *Proceedings of the International Conference on Ubiquitous Healthcare (uHealthcare)*, Korea, 2008.

Appendix A: Probabilistic Inference with Parametric-based Approach

In statistics, the probabilistic inference refers to a method of assigning a probability model to an event throughout empirical data. Also, it concerns about the probability computations of practical interest such as Bayesian estimation or searching for the mode of a specified probability distribution. In this appendix, we introduce probabilistic inference and its roles in computer vision, in which human pose and activity recognition is one of practical applications. Since the thesis focus is centered around the parametric-based approach, more essentials of probabilistic inference using this approach are overviewed accordingly.

A.1 Probabilistic Inference and Computer Vision

The utilizations of probabilistic inference in machine learning have been an active research area for last decades and provided many efficient methods to tackle various problems in fields including computer vision. Here, the probabilistic inference refers to the meaning of fitting a probability distribution to real data and performing further calculations on the data using the modeled distribution and probability rules. Meanwhile, the task of computer vision concerns about designing visual functions for a computer or machine such as recognizing objects, tracking objects' motions, or more complicated, understanding the context of scenes, from the images or video acquired by

digital cameras. Usually, computer vision plays fundamental roles in such areas as robots, remote sensing, or artificial intelligence. However, solving a practical problem in computer vision is always challenging, because besides the uncertainty of visual data in images and video, the information of an object in images cannot be easily well understood from the local features of isolated pixels. Many global factors should be taken into account in order to acquire the visual information of an object in image, e.g., the consistent appearances of objects with surrounding scenes [123]. Therefore, a general framework based on probabilistic inference to handle the uncertainty of visual data and to integrate different kinds of global information to address the problems of computer vision is receiving a lot of attention from research communities.

However, whether applying probabilistic inference is an appropriate choice to address a problem of computer vision. As being known, the developments of computer vision have been closely related to our understanding about how the brain works and what is the visual processing of the brain. So far, many studies in physiology and psychophysics have been attempted to discover the working mechanism of the human brain. Numerous biological evidences discovered by prominent researches suggested the new way of using the Bayesian probability theories to explain undiscovered functions of the brain [16, 49, 79]. In more details [73], in perception, the brain represents the sensory inputs (e.g., auditory and visual data) in its neural circuits to maximize the compatible probability of the stored information with the sensations. In decision making, the brain combines the new sensory information about the world with the learned experiments from the past (the prior) to give an action (i.e., send controlling signals to motor neurons to perform muscle contractions). After an activity is performed, the new knowledge, the likelihood evaluating the performance of the motor tasks with the observed sensations, might be integrated with the past prior to update the learned experiments (the posterior or the new prior) in a fashion formulated by Bayesian statistics. In the similar way, the visual processing could be expressed as estimating the posterior probability of visual features such as categorization, location, or texture of an object, given the pictures

observed though the human eyes [115, 116].

Regarding the knowledge about the visual processing inside the human brain, a task of computer vision is conveniently described by a probabilistic inference of estimating the posterior distribution $P(\vartheta|I)$ of visual features ϑ given an image I [113]. Using the Bayesian theorem, the posterior $P(\vartheta|I)$ is approximated by the product of the prior $P(\vartheta)$ and the likelihood $P(I|\vartheta)$, $P(\vartheta|I) \propto P(I|\vartheta)P(\vartheta)$, where the prior $P(\vartheta)$ presents the probability distributions of the visual features and the likelihood $P(I|\vartheta)$ evaluates the compatibility between the image and the visual features. From this equation, we can calculate the posterior $P(\vartheta|I)$ to update information about the visual features ϑ. Additionally, the optimal values ϑ^* of the random variable ϑ that maximizes the posterior distribution are correspondent to the visual features best compatible with the observed image. It is obviously that the established framework is based on probability to cope the uncertainty of input data and to deal with the object's appearances with varying in locations, shapes, and poses. Also, the framework is general to combine many global cues of visual features. However, in practice, the probabilistic model for a particular problem might be complicated and composed of a thousand of random variables. For such models, even a problem in computer vision is already converted to a problem in probability, the computations related to this probability are somehow intractable. Currently, there have been two approaches introduced to address complex probability inference in statistics and extended to be used in machine learning (and also computer vision). The first is namely the nonparametric-based approach and the other the parametric-based approach.

- *Nonparametric-based approach*: The underlying concept behind the nonparametric-based approach is that the random distribution of a complicated probability density function is represented by a set of sampled values (point mass or *"particle"*) of the random variables. In the situation that the random variable is uniformly sampled, each sample is assigned a

weight corresponding to its probability density at the sampling value. Meanwhile, if all of the samples have the same weight, the appearance frequency of each sample is proportional to the sample's probability. A set of samples is used to replace the associated probability for the further calculations on this probability.

- *Parametric-based approach*: The parametric-based approach expresses a particular problem in statistics as a solution of an analytical problem. Each probability needs to be parameterized with a specific probability density function and mathematical tools are applied to address the computations related to such probability distributions. Also, for the optimization problems in which the exact analytical solution does not exist, the problems could be mitigated by various ways including divide-and-conquer strategies and approximations.

The techniques, which have been applied throughout the thesis, are based on the *parametric-based* approach. For a defined probability distribution $p(z_1, z_2, ..., z_n)$, the three fundamental problems of probabilistic parametric inference [138] we need to address are

- *Marginal computation*: Estimating the marginal distribution of a joint density function p over a subset of random variables $X_s = \{z_{s_1}, z_{s_2}, ..., z_{s_l}\}$, where $X_s \subset \{z_1, z_2, ..., z_n\}$. The marginal distribution of continuous variables is directly calculated by integrating over a subset of random variables,

$$p(X) = \int p(X, Y) dY. \tag{A.1}$$

  Considering the probabilities with respect to discrete variables, the summarization is used to replace the integration in equation (A.1).

- *Optimization*: This is the task of finding the optimal values $(\hat{z}_1, \hat{z}_2, ..., \hat{z}_n)$ of $(z_1, z_2, ..., z_n)$ to maximize $p(z_1, z_2, ..., z_n)$, or equal to estimating the mode of a probability distribution.

- *Bayesian estimation*: The Bayesian estimation can be derived from the marginal computation: The estimations of the posterior distribution $p(\vartheta|X)$ are obtained by computing the joint distribution of $p(\vartheta, X)$, computing the marginal distribution of $p(X)$, and then applying an equation $p(\vartheta|X) = p(\vartheta, X)/p(X)$.

In order to compute more complicated inference of multivariate distributions, a graphical model is used to illustrate the conditional relationship among variables. Appropriate methods are proposed to perform the inference, dependent on the structure and complexity of a probability distribution.

A.2 Graphical Models of Probabilistic Distributions

In practice, the distribution formulated for a task of computer vision might be complicated when it contains a thousand of random variables. For example, the posterior $p(\vartheta|I)$ of the visual features given an image I can be completely expressed by $p(\vartheta_1^1, \vartheta_2^1, ..., \vartheta_m^n|I_1, I_2, ..., I_k)$, where $\vartheta_1^1, \vartheta_2^1, ..., \vartheta_m^n$ are various parameters including the location, shape, or name of n objects appearing in the image and $I_1, I_2, ..., I_k$ are the color of the k pixels in the image. Obviously, the calculations over such a joint density function are increased exponentially with the number of random variables. Thus, the such tasks as finding a maximum of the distribution or computing the marginal distribution always seem intractable. A general framework to address these difficulties is to construct a graphical model to analyze local relationships among random variables [18]. A joint distribution of all random variables is thereby factorized into the products of density distribution containing a subset of random variables. Based on this framework, we can eliminate computations that makes the use of probabilistic approach in computer vision become feasible. As follows, two kinds of graphical models using directed graph and undirected graphs are introduced correspondingly.

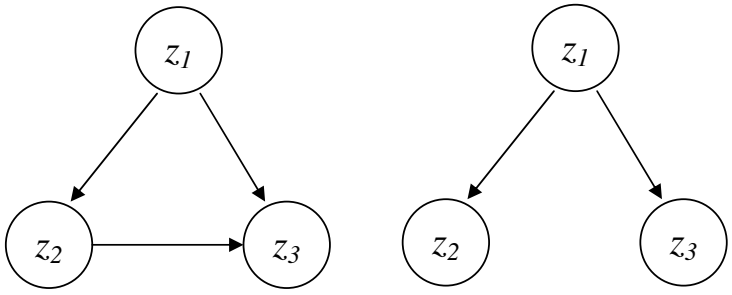
Model probabilistic distributions using directed graphs

Given a joint distribution, $p(z_1, z_2, ..., z_n)$, a directed graph of random variables, called *Bayesian network* [59, 92, 121], containing a set of nodes $\{z_1, z_2, ..., z_n\}$ as random variables and a set of directed edges is used to model this distribution. The edges connecting two nodes are associated by the relationship of conditional probabilities. For example, suppose we have the conditional probability $p(z_3|z_1, z_2)$, then the two directed edges $z_1 \rightarrow z_3$ and $z_2 \rightarrow z_3$ are included into a graph. Because a joint distribution has another presentation as a product of conditional distributions,

$$p(z_1, z_2, ..., z_n) = p(z_1)p(z_2|z_1)...p(z_n|z_1, ..., z_{n-1}) \tag{A.2}$$

there always exist a directed graph with full connections presenting a joint distribution. We give a particular example with a joint distribution $p(z_1, z_2, z_3)$ as depicted by Fig. A.1

The method to reduce the complexity of a directed graph is related to the concept of *condi-*



(a) (b)

Figure A.1: A directed graph used to describe a probability with conditional relationship. (a) A graph with full connections. (b) Using conditional independence to remove an edge.

tional independence. Considering the previous example with $p(z_1, z_2, z_3) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2)$,

the edge between z_1 and z_3 is eliminated if z_3 does not depend on z_2 given z_1

$$p(z_3|z_1, z_2) = p(z_3|z_1), \tag{A.3}$$

or in the other way

$$p(z_3, z_2|z_1) = p(z_3|z_1, z_2)p(z_2|z_1) = p(z_3|z_1)p(z_2|z_1). \tag{A.4}$$

The expression of equation (A.4) reveals the fact that z_3 and z_2 are conditionally independent given z_1. In practice, usually each random variable z_i only locally depends on a subset of random variables $\zeta(z_i)$. We are therefore able to apply the conditional independence of random variables to simplify the structure of the modeled graph, correspondingly reducing the computations based on the associated distribution,

$$p(z_1, z_2, ..., z_n) = \prod_{i=1}^{n} p(z_i|\zeta(z_i)). \tag{A.5}$$

Fig. A.2 shows an example of pruning out some edges of the full connected graph, resulting a simplification of a joint distribution:

$$p(z_1, z_2, ..., z_6) = p(v_1)p(v_2|v_1)p(v_3|v_1)p(v_4|v_1)p(v_5|v_2, v_3)p(v_6|v_3). \tag{A.6}$$

Model probabilistic distributions using undirected graphs

In another method, an undirected graph can also be used to represent a joint density function $p(z_1, z_2, ..., z_n)$ [18, 138]. An undirected graph contains a set of cliques where each cliques s is constructed by a full connected subset of nodes $\mathcal{Z}_s = \{z_i \in s\}$ and is associate with one function $\varphi_(\mathcal{Z}_s)$ getting a non-negative value. The factorization of the joint density function $p(z_1, z_2, ..., z_n)$ is given by

$$p(z_1, z_2, ..., z_n) = \frac{1}{Z} \prod_{s \subset S} \varphi_s(\mathcal{Z}_s) \tag{A.7}$$

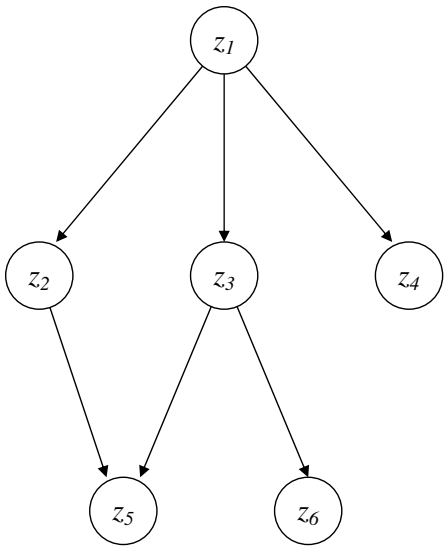


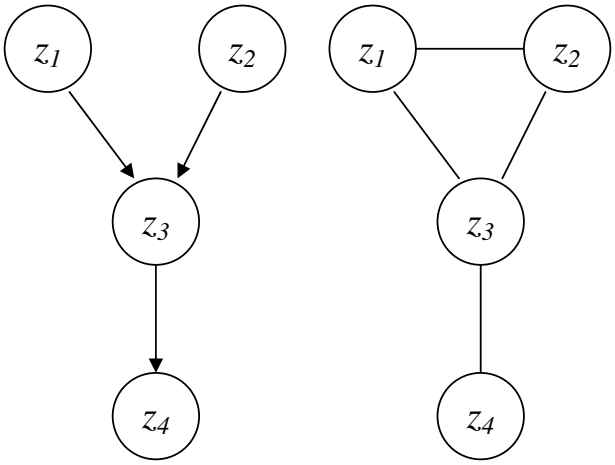
Figure A.2: A complicated distribution modeled by a directed graph after simplified.

where S is the number of cliques and Z is a normalized constant. Because there are no direct relations are illustrated in this presentation, the probabilistic distributions are said to be modeled by an undirected graph. Generally, it is possible to convert a directed graph into an undirected graph. Given an example in Fig. A.3, the directed graph depicted in Fig. A.3(a) corresponds to a joint density function

$$p(z_1, z_2, z_3, z_4) = p(z_3|z_1, z_2)p(z_1)p(z_2)p(z_4|z_3). \tag{A.8}$$

Suppose an undirected graph is given in Fig. A.3(b). Then the joint density function $p(z_1, z_2, z_3, z_4)$ can be factorized into the products of two functions $\varphi_{1,2,3}(z_1, z_2, z_3)$ and $\varphi_{3,4}(z_3, z_4)$ defined on the two cliques $\{z_1, z_2, z_3\}$ and $\{z_3, z_4\}$

$$p(z_1, z_2, z_3, z_4) = \varphi_{1,2,3}(z_1, z_2, z_3)\varphi_{3,4}(z_3, z_4) \tag{A.9}$$



(a) (b)

Figure A.3: The differences between a directed graph and an undirected graph when we model the same distribution. (a) A directed graph. (b) An undirected graph.

where $\varphi_{1,2,3}(z_1, z_2, z_3) = p(z_3|z_1, z_2)p(z_1)p(z_2)$ and $\varphi_{3,4}(z_3, z_4) = p(z_4|z_3)$.

In Fig. A.4, an undirected graph, pairwise Markov random fields (MRF) [70] arising in many applications such as image denoising, stereo matching, image segmentation, etc. is introduced. The density function described by this graph is expressed by

$$p(z|x) = \frac{1}{Z} \prod_{(i,j)} \varphi_{i,j}(z_i, z_j) \prod_i \varphi_i(z_i, x_i) \tag{A.10}$$

where Z is a normalized constant.

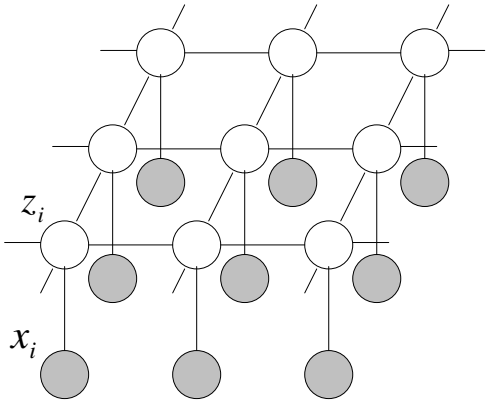


Figure A.4: Markov random fields.

A.3 Probabilistic Parametric Inference on Probabilistic Graphical Models

There exist analytical and numerical methods for the probabilistic inference of simple distributions. For instance, the problem of maximizing or minimizing a distribution is directly solved using available optimization methods. Some typical methods of estimating optimal values of a function often used in various areas include gradient ascent or descent method, conjugate gradient method, Newton's method, etc. However, the direct calculation of probabilistic inference is only trivial with simple distributions consisting of a small number of variables. The integration of a continuous function over numerous variables is a very difficult task. This is also true with finding an optimal value of multivariate functions. Based on the graphical structure modeled for a probabilistic problem, exact methods (see Appendix B) or appropriate methods (see Appendix C) are appropriately suggested to address the probabilistic inference for the particular distribution.

Appendix B: Exact Probabilistic Inference for HMMs and Kalman Filter

In the special but common case in which a graphical model of a distribution has the form of tree, a graph without cycles, there exist exact methods for addressing the inference problems of this distribution. The underlying idea of such methods is that we divide a complicated problem into subproblems and combine the results of involved subproblems to get an overall solution. A sample probability distribution described by a tree is given in Fig. B.1. It is straightforward that the random variables at the depth k of a tree only conditionally depend on the random variables at the depth $k-1$. Let denote $Z_k = \{z_i | z_i \in \text{depth } k\}$ be a set of random variables belonging to the depth k. The expression formulated this conditional relationship is given by

$$p(Z_k|Z_1, Z_2, ..., Z_{k-1}) = p(Z_k|Z_{k-1}). \tag{B.1}$$

Therefore, a joint distribution $p(Z_1, Z_2, ..., Z_k)$ can be computed by

$$\begin{aligned} p(Z_1, Z_2, ..., Z_k) &= p(Z_k|Z_1, Z_2, ..., Z_{k-1})p(Z_1, Z_2, ..., Z_{k-1}) \\ &= p(Z_k|Z_{k-1})p(Z_1, Z_2, ..., Z_{k-1}). \end{aligned} \tag{B.2}$$

Equation (B.2) shows that the inference of the joint distribution $p(Z_1, Z_2, ..., Z_k)$ can be relied on the results of previous inference of $p(Z_1, Z_2, ..., Z_{k-1})$. This leads to a recursive strategy to get

an exact inference of a hierarchy model. To make it clear, we illustrate the method of computing a marginal probability distribution over a tree. Assume that we want to estimate the marginal distribution over all of the depth levels $1, 2, ..., k$

$$p_k(X) = \int p(Z_1, Z_2, ..., Z_k, X) dZ_1 dZ_2 ... dZ_k. \tag{B.3}$$

We now substitute equation (B.2) into (B.3), giving

$$\begin{aligned} p_k(X) &= \int p(Z_1, Z_2, ..., Z_{k-1}, X) p(Z_k|Z_{k-1}) dZ_1 dZ_2 ... dZ_k \\ &= \int p(Z_k|Z_{k-1}) \left(\int p(Z_1, Z_2, ..., Z_{k-1}, X) dZ_1 dZ_2 ... dZ_{k-1} \right) dZ_k \\ &= \int p(Z_k|Z_{k-1}) p_{k-1}(X) dZ_k. \end{aligned} \tag{B.4}$$

Clearly, this is a nice recursive formulation to compute the marginal of $p_k(X)$ over $Z_1, Z_2, ..., Z_k$ throughout the marginal of $p_{k-1}(X)$ over $Z_1, Z_2, ..., Z_{k-1}$. Repeating this operation beginning from $p_0(X)$ along with increasing the depth level, we obtain the marginal distributions for all random variables. The similar mechanism of formulating a complicated inference by a recursive sequence of simpler inferences is also used to maximize or minimize a multivariate distribution with tree-structured graphical models.

The two instances of the described graphical models involves hidden Markov model (HMM) and Kalman filter. The HMM has been popular for the use in discovering a pattern from sequence data such as speech, video, genes, etc. The Kalman filter has been adaptable in a linear dynamical system (LDS) in which we want to measure the a value of an unknown quality z from the observation x acquired by a noisy sensor. The HMM and Kalman filter share the same hierarchy model as shown by Fig B.2.

A density distribution with respect to random variables $z_0, z_2, .., z_n$ and $x_0, x_2, ..., x_n$ is ex-

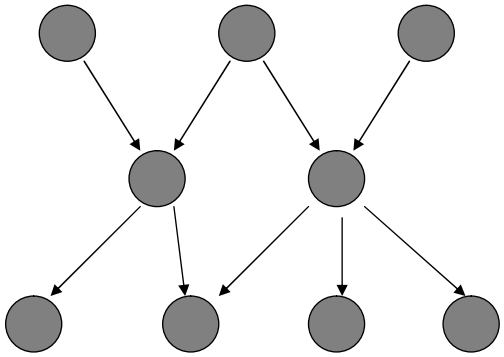


Figure B.1: A tree-structured graphical model.

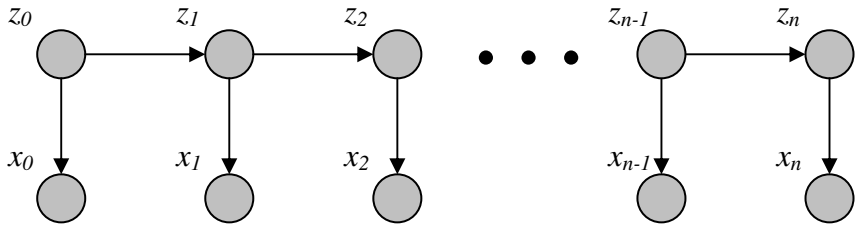


Figure B.2: A graphical model of HMM and Kalman filter.

pressed by

$$p(Z, X) = p(z_0)p(x_0|z_0)\prod_{i=1}^{n} p(z_i|z_{i-1})\prod_{i=0}^{n} p(x_i|z_i). \tag{B.5}$$

Here, $Z = \{z_0, z_2, ..., z_n\}$ is a set of state or hidden variables and $X = \{x_0, x_1, ..., x_n\}$ is a set of observed variables. However, the HMM is defined with discrete state variables, meanwhile the Kalman filter replies on a distribution of continuous random variables.

Mathematically, a multinomial distribution is used to formulate the conditional probability of a HMM with the 1-of-K presentation for a state variable (each random variable z_i is presented by

a set of discrete elements $z_i^1, z_i^2, ..., z_i^k$, in which just one element z_i^k gets a value of one, while the others are set to zeros),

$$p(z_i|z_{i-1}) = \prod_{k=1}^{K}\prod_{k=1}^{K} S_{jk}^{z_{i-1}^k z_i^k}$$
$$p(z_0) = \prod_{k=1}^{K} \pi_k^{z_0^k} \tag{B.6}$$

where S_{jk} is known as a matrix of transition probability and π_k is an initial probability. Alternatively, we express the conditional probability $p(z_i|z_{i-1})$ using a Gaussian distribution in the case of the Kalman filter

$$p(z_i|z_{i-1}) = \mathcal{N}(z_i|Az_{i-1}, \Sigma_i)$$
$$p(z_0) = \mathcal{N}(z_0|m_0, \Sigma_0) \tag{B.7}$$

where Σ_i is a covariance matrix of Gaussian noise and m_0 is an initial values of z. In order to update the whole conditional probability $p(z_i|z_{i-1}, z_{i-2}, ..., z_0)$, we utilize the two equations as follows to calculate the Gaussian distribution of y and the conditional distribution of x given y

$$p(y) = \mathcal{N}(y|A\mu + b, \Sigma_2^{-1} + A\Sigma_1^{-1}A^T) \tag{B.8}$$
$$p(x|y) = \mathcal{N}(x|\Sigma\{A^T\Sigma_2(y-b) + \Sigma_1\mu\}, \Sigma) \tag{B.9}$$
$$\Sigma = (\Sigma_1 + A^T\Sigma_2 A)^{-1} \tag{B.10}$$

where we already have an assumption

$$p(x) = \mathcal{N}(x|\mu, \Sigma_1^{-1}) \tag{B.11}$$
$$p(y|x) = \mathcal{N}(y|Ax + b, \Sigma_2^{-1}) \tag{B.12}$$

More essentials of HMM are available in [67, 74] and of Kalman filter in [68, 149] respectively.

Appendix C: Variational Inference with Expectation Maximization and Variational Expectation Maximization

For a probability distribution defined on a complicated graphical model which does not have a tree-structure, it remains challenging to perform exact parametric inference. Therefore, numerous approximate approaches have been concerned so far to estimate just a closed optimal solution for such inference. We mainly concentrate on the variational method and illustrate the applications of this method for human pose estimation. First of all, we introduce the basic concepts related to variational methods. Actually, the original of the variational method does not refer to the meaning of approximations but to the calculus of variations, which is used to find functions to minimize or maximize the value of quantities defined over these functions. The Euler-Lagrange equation [47, 53] is the best known outcomes of this approach, which has yielded successes in many areas of mathematics and physics. In the following, we draw our focus to the use of the variational approach to address probabilistic inference with the EM algorithm and its extension with the VEM algorithm.

C.1 Expectation Maximization

In general, we may need to maximize the likelihood function containing hidden variables

$$\text{argmax}_{\theta^*} p(X|\theta) = \int p(X, Z|\theta) dZ. \tag{C.1}$$

However, the direct calculation of this optimization problem is somehow very complicated, meanwhile the optimization on the complete-likelihood $p(X, Z|\theta)$ may get easier suggesting that we can transform an optimization of $p(X|\theta)$ into a less complex optimization of $p(X, Z|\theta)$. The EM algorithm is developed based on this underlying idea [39, 85]. With a variation $f(Z)$ defined over the latent variables and with any choice of $f(Z)$, we always have

$$\begin{aligned} \ln p(X|\theta) &= \int f(Z) \ln \frac{p(X, Z|\theta)}{f(Z)} dZ + D_{KL}(f||p) \\ &= E_f[p(X, Z|\theta)] + D_{KL}(f||p) - \int f(Z) \ln f(Z) dZ \end{aligned} \tag{C.2}$$

where $D_{KL}(f||p) = -\int f(Z) \ln \frac{p(Z|X,\theta)}{f(Z)} dZ$ is the Kullback-Leibler distance between f and $p(Z|X, \theta)$, that is always non-negative $D_{KL}(f||p) \geq 0$. The equality happens if and only if $f(Z) = p(Z|X, \theta)$. Regarding the non-negative properties of the Kullback-Leibler distance and (C.2), the EM algorithm proposed to solve the problem in (C.1) is expressed by an iterative procedure whose each iteration consists of the following two main steps:

i) E-step: Assume that the current value of θ is $\theta_{\textbf{old}}$. The E-step evaluates the analytical expression of the posterior distribution $f(V)$ as

$$f(Z) = p(Z|X, \theta_{\textbf{old}}). \tag{C.3}$$

ii) M-step: The M-step maximizes

$$E_{f_{\textbf{old}}(Z)}[\log p(X, Z|\theta)] \tag{C.4}$$

with respect to θ where $f_{\textbf{old}}(Z)$ is found from the previous E-step.

The algorithm is iterated until it converges to a stable value of θ.

C.2 Variational Expectation Maximization

Previously, the EM algorithm has been introduced to estimate the maximum values of a distribution $p(X)$ from the joint distribution $p(X, Z)$ containing latent variables Z. However, in particular it might become difficult to get the analytical expression of the posterior distribution $p(Z|X)$ in the E-step. The way to reduce the complexity of $p(Z|X)$ is to approximate it by close and simpler distributions such that we can perform the inference on the replaced distributions. As already known, it is able to formulate the statistical model of a complicated distribution in terms of a graph, namely a graphical model. The task of inference can be analytically performed on a distribution with a tree-structure. Therefore, with the graphical model of a complicated distribution, we can remove specified edges to build a spanning tree of this graph (The a spanning tree is the best reconstruction from a graph with minimum eliminations of edges). A distribution described by a spanning tree seems to achieve good appropriate approximation of the targeted distribution.

Alternatively, the mean-field algorithm [97] demonstrates a simpler method by partitioning a distribution into the products of distributions independent over disjoint subset of random variables. Here, we illustrate the way of formulating the variational method in this way. Let denote the original distribution be $p(Z|X)$ and the approximation of this distribution be $f(Z)$. We assume that $f(Z)$ is independent over a subset of random variables Z_i where $i = 1, 2, ..., N$,

$$f(Z) = \prod_{i=1}^{N} f_i(Z_i) \tag{C.5}$$

Here, Z_i is a disjoint subset of Z. There are no restrictions on the form of $f_i(Z_i)$, so that we need to select an appropriate formulation of f_i to make $f(Z)$ to be as close as possible to $p(Z|X)$ by minimizing the Kullback-Leibler distance $D_{KL}(f||p(Z|X))$ between them. Similar to (C.2), we

have

$$\ln p(X) = E_f[p(X,Z)] + D_{KL}(f||p(Z|X)) - \int f(Z)\ln f(Z)dZ. \tag{C.6}$$

Since the form of $p(X)$ does not depend on Z and is considered as a constant with respect to Z, minimizing $D_{KL}(f||p(Z|X))$ is correspondent to

$$\text{argmax}_f \mathcal{F}(f) = E_f[p(X,Z)] - \int f(Z)\ln f(Z)dZ. \tag{C.7}$$

We now substitute the explicit form of $f(Z)$ in (C.5) into $E_f[p(X,Z)]$ to get

$$\begin{aligned}
\mathcal{F}(f) &= \int \prod_j f_j(Z_j)\ln p(X,Z)dZ - \sum_j \int \prod_j \ln f_j(Z_j)dZ \\
&= \sum_j \left\{ \int f_j E_{i\neq j}[\ln p(X,Z)]dZ_j - \int f_j \ln f_j dZ_j \right\} + const \\
&= \sum_j -D_{KL}(f_j|\exp\{E_{i\neq j}[\ln p(X,Z)]\}) + const
\end{aligned} \tag{C.8}$$

where $E_{i\neq j}[\ln p(X,Z)] = \int \ln p(X,Z)\prod_{i\neq j} f_i(Z_i)dZ_i$. It is obviously that maximizing the value of $\mathcal{F}(f)$ is actually equal to minimizing $D_{KL}(f_j|\exp\{E_{i\neq j}[\ln p(X,Z)]\})$, satisfied when

$$f_j(Z_j) = \xi_j \exp\{E_{i\neq j}[\ln p(X,Z)]\} \tag{C.9}$$

where ξ_j is a normalize constant that makes sure $\int f_j(Z_j)dZ_j = 1$. An equation of $f(Z)$ formulated over a set of equation $f_j(Z_j)$ defined by (C.9) is the best approximation of $p(Z|X)$ in term of minimizing the Kullback-Leibler distance.

Similar to the EM algorithm, the VEM algorithm is proposed with the two main steps:

i) Variational E-step (VE-step): Assuming that the current value of θ is $\theta_{\textbf{old}}$, the VE-step approximates the posterior $p(Z|X,\theta_{\textbf{old}})$ by a distribution $f^*(V)$ such that the Kullback-Leibler distance between them is minimize

$$f^*(Z) = \text{argmin}_f D_{KL}(f||p(Z|X,\theta_{\textbf{old}})). \tag{C.10}$$

ii) M-step: The M-step maximizes

$$E_{f^*(Z)}[\log p(X, Z|\theta)] \tag{C.11}$$

with respect to θ where $f^*(Z)$ is found from the previous E-step.

The algorithm is iterated until it converges to a stable value of θ.

Appendix D: Locating the Nearest Point in an Ellipsoid Surface to a Given Point

In Fig. 3.5, the transformation of X_0 into the local coordinate system (x', y', z') attached to the ellipsoid is obtained as

$$[x'_0, y'_0, z'_0, 1]^T = \mathbf{SQ}_{\vartheta}[x_0, y_0, z_0, 1]^T. \tag{D.1}$$

In the 2-D coordinate system of the plane P (the origin of the plane P lies at the centroid of the ellipsoid), these coordinates are converted to $(\sqrt{x'^2_0 + y'^2_0}, z'_0)$. The intersection between the plane P and the ellipsoid will be an ellipse with the major axis c and the minor axis a. Hence, the nearest point X_t belonging to the ellipsoid surface of X_0 in the plane P has the 2-D coordinate (u, v) as the roots of the equation

$$\begin{aligned} f(u,v) &= \frac{u^2}{a^2} + \frac{v^2}{c^2} - 1 = 0 \\ (\sqrt{x'^2_0 + y'^2_0} - u)\frac{\partial f(u,v)}{\partial v} &= (z'_0 - v)\frac{\partial f(u,v)}{\partial u}. \end{aligned} \tag{D.2}$$

This equation can be converted to a fourth-degree polynomial equation to find u and v. The coordinate of X_t in (x', y', z') is given by

$$x'_t = u\frac{x'_0}{\sqrt{x'^2_0 + y'^2_0}}, y'_t = u\frac{y'_0}{\sqrt{x'^2_0 + y'^2_0}}, z'_t = v. \tag{D.3}$$

We expand the updated rules for computing $X_t(x'_t, y'_t, z'_t)$ when X_0 moves to $X_0 + dX_0$ in the local coordinate system (x', y', z'). Let $k = a^2/c^2$ be a constant. Let

$$\begin{aligned} u &= \sqrt{x_t'^2 + y_t'^2}, v = z'_t \\ r'_0 &= \sqrt{x_0'^2 + y_0'^2} \\ \cos\gamma &= x'_0/r'_0, \sin\gamma = y'_0/r'_0. \end{aligned} \tag{D.4}$$

The new value of $X_t(x'_t, y'_t, z'_t)$ corresponding to $X_0 + dX_0$ = $(x'_0 + dx'_0, y'_0 + dy'_0, z'_0 + dz'_0)$ in the local coordinate system (x', y', z') is computed by

$$\begin{aligned} \xi &= \frac{kv(\cos\gamma dx'_0 + \sin\gamma dy'_0) - udz'_0}{(1-k)(kv^2 - u^2) - k(vz'_0 + ur'_0)} \\ u &= u - kv\xi, v = v + u\xi \\ x'_t &= u\cos\gamma, y'_t = u\sin\gamma, z'_t = v. \end{aligned} \tag{D.5}$$

Note that when the point moves from outside into the ellipsoid or vice versa, x'_t, y'_t, and z'_t need to be recomputed from (D.2) and (D.3).

Transforming $X_t(x'_t, y'_t, z'_t)$ back to the global coordinate system (x, y, z), the coordinate of X_t is given by

$$[x_t, y_t, z_t, 1]^T = \mathbf{Q}_\vartheta^{-1}\mathbf{S}^{-1}[x'_t, y'_t, z'_t, 1]^T.$$

Appendix E: Computation of the Jacobian Matrix for the Inverse Kinematic Problem

In this appendix, we focus on the computation of the Jacobian matrix **J** of $Z_i(\vartheta)^\varepsilon$ with respect to ϑ. Assuming that the ellipsoid ε depends on the n_ε parameters $\vartheta_1, \vartheta_2, ..., \vartheta_{n_\varepsilon}$, $[Z_i(\vartheta)^\varepsilon, 1]^T$ must satisfy equation (3.3) with $\vartheta = (\vartheta_1, \vartheta_2, ..., \vartheta_{n_\varepsilon})$. Because $Z_i(\vartheta)^\varepsilon$ belongs to an ellipsoid surface in the global coordinate system, we apply a series of transformations to $Z_i(\vartheta)^\varepsilon$ to get one point $Z0_i^\varepsilon$, independent of ϑ, lying in an ellipsoid surface in the local coordinate system

$$\begin{aligned} &\mathbf{S}\mathbf{Q}_{n_\varepsilon}(\vartheta_{n_\varepsilon})\mathbf{Q}_{n_\varepsilon-1}(\vartheta_{n_\varepsilon-1})...\mathbf{Q}_1(\vartheta_1)[Z_i(\vartheta)^\varepsilon, 1]^T = [Z0_i{}^\varepsilon, 1]^T \\ &\text{or } [Z_i(\vartheta)^\varepsilon, 1]^T = \mathbf{Q}_1(\vartheta_1)^{-1}\mathbf{Q}_2(\vartheta_2)^{-1}...\mathbf{Q}_{n_\varepsilon}(\vartheta_{n_\varepsilon})^{-1}\mathbf{S}^{-1}[Z0_i{}^\varepsilon, 1]^T. \end{aligned} \tag{E.1}$$

The Jacobian matrix **J** consists of n_ε columns, where each column i, $\partial Z_i(\vartheta)^\varepsilon / \partial \vartheta_i$ is given by

$$
\begin{aligned}
&[\frac{\partial Z_i(\vartheta)^\varepsilon}{\partial \vartheta_i}, 0]^T = \mathbf{Q}_1(\vartheta_1)^{-1}\mathbf{Q}_2(\vartheta_2)^{-1}...\frac{\partial \mathbf{Q}_i(\vartheta_i)^{-1}}{\partial \vartheta_i} \\
&\mathbf{Q}_{i+1}(\vartheta_{i+1})^{-1}...\mathbf{Q}_{n_\varepsilon}(\vartheta_{n_\varepsilon})^{-1}\mathbf{S}^{-1}[Z0_i{}^\varepsilon, 1]^T \\
&= \mathbf{Q}_1(\vartheta_1)^{-1}\mathbf{Q}_2(\vartheta_2)^{-1}...\frac{\partial \mathbf{Q}_i(\vartheta_i)^{-1}}{\partial \vartheta_i} \\
&\mathbf{Q}_i(\vartheta_i)...\mathbf{Q}_2(\vartheta_2)\mathbf{Q}_1(\vartheta_1)[Z_i(\vartheta)^\varepsilon, 1]^T. \qquad \text{(E.2)}
\end{aligned}
$$

References

[1] Fujji Finepix Real 3D camera http://www.fujifilm.com/products/3d/camera/.

[2] Gypsy motion capture system http://www.metamotion.com/gypsy/gypsy-motion-capture-system.htm.

[3] Minoru stereo camera http://www.minoru3d.com.

[4] Motion capture systems from Vicon http://www.vicon.com/.

[5] MVN-inertial motion capture http://www.xsens.com/en/general/mvn/.

[6] Stereo camera Bumblebee 2.0 http://www.ptgrey.com/products/stereo.asp.

[7] M. F. Abdelkader, A. K. Roy-Chowdhury, R. Chellappa, and U. Akdemir. Activity representation using 3D shape models. *EURASIP Journal on Image and Video Processing*, 2008(347050):16–pages, 2008.

[8] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.

[9] B. Allen, B. Curless, and Z. Popovic. Articulated body deformation from range scan data. *ACM Transactions on Graphics*, 21(3):612–619, 2002.

[10] B. Allen, B. Curless, and Z. Popovic. The space of all body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics*, 22(3):587–594, 2003.

[11] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005.

[12] H. Attias. A variational Bayesian framework for graphical models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 209–215, Denver, CO, USA, December 2000.

[13] C. Barron, Ioannis, and A. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding*, 81(3):269–284, 2001.

[14] C. Barron and I. A. Kakadiaris. Estimating anthropometry and pose from a single image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 669–676, Hilton Head, SC, USA, June 2000.

[15] C. Barron and I. A. Kakadiaris. On the improvement of anthropometry and pose estimation from a single uncalibrated image. *Machine Vision and Applications*, 14(4):229–236, 2003.

[16] J. Beck, W. J. Ma, R. Kiani, T. Hanks, A. K. Churchland, L. Roitman, M. N. Shadlen, P. Latham, and A. Pouget. Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.

[17] P. N. Belhumeur, J. Espanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[18] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[19] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[20] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[21] G. J. Brostow, I. Essa, D. Steedly, and V. Kwatra. Novel skeletal representation for articulated creatures. In T. Pajdla and J. Matas, editors, *Proceedings of European Conference on Computer Vision*, pages 11–14, Prague, Czech Republic, May 2004.

[22] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Proceedings of European Conference on Computer Vision*, pages 326–339, Heraklion, Crete, Greece, September 2010.

[23] C. Cagniartand, E. Boyer, and S. Ilic. Iterative deformable surface tracking in multi-view setups. In *Proceedings of the Fifth International Symposium on 3D Data Processing, Visualization and Transmission*, Paris, France, May 2010.

[24] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. In *Proceedings of the IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*, pages 263–270, Kauai, HI, USA, December 2001.

[25] J. Cech and R. Sara. Efficient sampling of disparity space for fast and accurate matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, US, June 2007.

[26] T. H. Chalidabhongse, K. Kim, D. Harwood, and L. Davis. A perturbation method for evaluating background subtraction algorithms. In *Proceedings of Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 15–16, Beijing, China, October 2005.

[27] I. Chang and S. Y. Lin. 3D human motion tracking based on a progressive particle filter. *Journal of Pattern Recognition*, 43(10):3612–3635, 2010.

[28] V. P. R. Chellappa. View independent human body pose estimation from a single perspective image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 16–22, Washington, DC, USA, June 2004.

[29] C. Chen, Y. Yang, F. Nie, and J.-M. Odozez. 3D human pose recovery from image by efficient visual feature selection. *Computer Vision and Image Understanding*, DOI: 10.1016/j.cviu.2010.11.007, 2010.

[30] G. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette for articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 16–22, Madison, Wisconsin, USA, June 2003.

[31] C.-W. Chu, O. C. Jenkins, and M. J. Mataric. Markerless kinematic model and motion capture from volume sequences. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 475–482, Beijing, China, October 2003.

[32] D. Comaniciu and P. Meer. Meanshift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[33] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[34] C. O. Conaire, N. E. O'Connor, and A. F. Smeaton. Detector adaption by maximising agreement between independent data sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, Minneapolis, MN, USA, June 2007.

[35] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Boca Raton, Florida: Chapman and Hall, 2001.

[36] M. V. d. Bergh, E. Koller-Meier, and L. V. Gool. Real-time body pose recognition using 2D or 3D Haarlets. *International Journal of Computer Vision*, 83(1):72–84, 2009.

[37] B. J. d'Auriol, T. Nguyen, T. Pham, S. Lee, and Y.-K. Lee. Viewer perception of superellipsoid-based accelerometer visualization techniques. In *Proceedings of The 2008 International Conference on Modeling, Simulation and Visualization Methods*, pages 129–135, Las Vegas, Nevada, USA, July 2006.

[38] D. Demirdjian. Combinining geometric- and view-based approaches for articulated pose estimation. In *Proceedings of the Eight European Conference on Computer Vision*, pages 183–194, Prague, Czech, May 2004.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from imcomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

[40] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61(2):185–205, 2005.

[41] K. I. Diamantaras and S. Y. Kung. *Principal Component Neural Networks: Theory and Applications*. Wiley, 1996.

[42] D. E. DiFranco, T.-J. Cham, and J. M. Rehg. Reconstruction of 3-D figure motion from 2-D correspondences. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 307–314, Kauai, HI, USA, December 2001.

[43] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[44] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *Proceedings of the International Conference On Computer Vision*, pages 315–320, Vancouver, Canada, July 2001.

[45] J. Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

[46] A. M. Elgammal and C. S. Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 681–688, Washington, DC, USA, June 2004.

[47] A. R. Forsyth. *Calculus of Variations*. New York: Dover, 1960.

[48] N. Friedman. The Bayesian structural EM algorithm. In *Proceedings of Conference on Uncertainty in Articial Intelligence*, pages 129–138, San Francisco, CA, USA, 1998.

[49] K. Friston. The free-energy principle: A unified brain theory? *Nature Neuroscience*, 11:1432 – 1438, 2006.

[50] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture: A multi-layer framework. *International Journal of Computer Vision*, 87(1-2):75–92, 2010.

[51] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, CA, USA, June 1996.

[52] Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Proceedings of Advances in Neural Information Processing Systems*, pages 449–455, Denver, CO, USA, December 2000.

[53] H. Goldstein. *Classical Mechanics*. Addison-Wesley, 1980.

[54] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

[55] D. Grest, J. Woetzel, and R. Koch. Nonlinear body pose estimation from depth images. *Lecture Notes in Computer Science*, 3663:285–292, 2005.

[56] A. Gupta, A. Mittal, and L. S. Davis. Constraint integration for efficient multiview pose estimation with self-occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):493–506, 2008.

[57] S. Hauberg and K. S. Pedersen. Predicting articulated human motion from spatial processes. *International Journal of Computer Vision*, DOI: 10.1007/s11263-011-0433-3, 2011.

[58] P. S. Heckbert. *Graphics Gems IV*. Academic Press, 1994.

[59] D. Heckerman, A. Mamdani, and M. P. Wellman. Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–68, 1995.

[60] H. Hirschmuller, P. R. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3):229–246, 2002.

[61] R. Horaud, M. Niskanen, G. Dewaele, and E. Boyer. Human motion tracking by registering an articulated surface to 3D points and normals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):158–163, 2009.

[62] N. R. Howe. Silhouette lookup for automatic pose tracking. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, page 15, Los Alamitos, CA, USA, June 2004.

[63] N. R. Howe. Flow lookup and biological motion perception. In *Proceedings of the Internation Conference on Image Processing*, pages 1168–1171, Genova, Italy, September 2005.

[64] G. Hua, M. Yang, and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 747–754, San Diego, CA, USA, June 2005.

[65] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

[66] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

[67] T. Kailath. A view of three decades of linear filtering theory. *IEEE Transactions on Informatin Theory*, 20(2):146–181, 1974.

[68] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the American Society for Mechanical Engineering, Series D, Journal of Basic Engineering*, 82:35–45, 1960.

[69] J. Karhunen, A. Cichocki, W. Kasprzak, and P. Pajunen. On neural blind separation with noise suppression and redundancy reduction. *International Journal of Neural Systems*, 8(2):219–237, 1997.

[70] R. Kindermann and L. J. Snell. *Markov random fields and their applications*. American Mathematical Society, 1980.

[71] O. D. King and D. A. Forsyth. How does CONDENSATION behave with a finite number of samples? In D. Vernon, editor, *Proceedings of European Conference on Computer Vision*, pages 695–709, Dublin, Ireland, June 2000.

[72] D. Knossow, R. Ronfard, and R. Horaud. Human motion tracking with a kinematic parameterization of extremal contours. *International Journal of Computer Vision*, 79(3):247–269, 2008.

[73] K. Kording. Decision theory: What should the nervous system do? *Science*, 318(5850):606–610, 2007.

[74] R. Lawrence and A. Rabiner. Tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[75] H. J. Lee and Z. Chen. Determination of 3D human body posture from a single view. *Computer Vision, Graphics and Image Processing*, 30(2):148–168, 1985.

[76] M. W. Lee and I. Cohen. A model-based approach for estimating human 3D poses in static images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):905–916, 2006.

[77] M. W. Lee, I. Cohen, and S. K. Jung. Particle filter with analytical inference for human body tracking. In *Proceedings of the Workshop on Motion and Video Computing*, pages 159–168, Orlando, FL, USA, June 2002.

[78] B. Luo and E. Hancock. Structural graph matching using the EM algorithm and singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1120–1136, 2001.

[79] W. J. Ma, J. Beck, P. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Science*, 9(5850):606–610, 2007.

[80] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In D. Vernon, editor, *Proceedings of European Conference on Computer Vision*, pages 3–19, Dublin, Ireland, June 2000.

[81] D. MacKay. Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge, 1997.

[82] D. Mateus, R. P. Horaud, D. Knossow, F. Cuzzolin, and E. Boyer. Articulated shape matching using laplacian eigenfunctions and unsupervised point registration. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, USA, June 2008.

[83] L. Maundermann, S. Corazza, and T. Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of Neuroengineering and Rehabilitation*, 3(6):185–205, 2006.

[84] J. M. McCarthy. *Introduction to Theoretical Kinematics*. Cambridge-MIT Press, 1990.

[85] G. J. McLachlan and T. Krishman. *The EM Algorithm and Its Extensions*. Wiley, 1997.

[86] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889, 2001.

[87] T. B. Moeslund, A. Hilton, and V. Krger. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.

[88] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006.

[89] G. Mori, X. Ren, A. A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 326–333, Washington, DC, USA, July 2004.

[90] K. Muhlmann, D. Maier, J. Hesser, and R. Manner. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision*, 47(1-3):79–88, 2002.

[91] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. Ann Arbor-CRC Press, 1994.

[92] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, NJ, 2004.

[93] F. Niu and M. Abdel-Mottaleb. View-invariant human activity recognition based on shape and motion features. In *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering*, pages 546–556, Miami, FL, USA, December 2004.

[94] F. Niu and M. Abdel-Mottaleb. HMM-based segmentation and recognition of human activities from video sequences. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 804–807, Amsterdam, Netherlands, July 2005.

[95] E. Oja. *Subspace Methods of Pattern Recognition*. Research Studies Press, England and Wiley USA, 1983.

[96] E.-J. Ong, A. S. Micilotta, R. Bowden, and A. Hilton. Viewpoint invariant exemplar-based 3D human tracking. *Computer Vision, Graphics and Image Processing*, 104(2-3):178–189, 2006.

[97] G. Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.

[98] S. I. Park and J. K. Hodgins. Capturing and animating skin deformation in human motion. *ACM Transactions on Graphics*, 25(3):881–889, 2006.

[99] K. Person. Onlines and planes of closest fit to systems of points in space. *Philosophical Magasize*, 2:559–572, 1901.

[100] P. Peursum, S. Venkatesh, and G. West. A study on smoothing for particle filtered 3D human body tracking. *International Journal of Computer Vision*, 87(1-2):53–74, 2010.

[101] R. Plankers and P. Fua. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding*, 81(3):285–302, 2001.

[102] R. Plankers and P. Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1182–1187, 2003.

[103] R. Pless. Image spaces and video trajectories. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1433–1440, Nice, France, October 2003.

[104] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1–2):4–18, 2007.

[105] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):65–81, 2007.

[106] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. A. Viola. Learning silhouette features for control of human motion. *ACM Transactions on Computer Graphics*, 24(4):1303–1331, 2005.

[107] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proceeding of the IEEE International Conference on Computer Vision*, volume 1, pages 824–831, Beijing, China, October 2005.

[108] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman filter: Particle Filters for Tracking Applications*. Artech House, Boston, London, 2004.

[109] T. J. Roberts, S. J. McKenna, and I. W. Ricketts. Human pose estimation using partial configurations and probabilistic regions. *International Journal of Computer Vision*, 73(3):285–306, 2007.

[110] R. Rosales and S. Sclaroff. Specialized mappings and the estimation of human body pose from a single image. In *Proceedings of the IEEE Workshop on Human Motion (HUMO)*, pages 19–24, Austin, TX, USA, December 2000.

[111] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 209(5500):2323 – 2326, 2000.

[112] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1709–1718, New York, NY, USA, June 2006.

[113] C. T. S. Chikkerur, T. Serre and T. Poggio. What and where: A Bayesian inference theory of attention. *Vision Research*, 55(22):2233–2247, 2010.

[114] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[115] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science*, 104(15):6424–6429, 2007.

[116] T. Serre and T. Poggio. A neuromorphic approach to computer vision. *Communications of the ACM*, 53(10):54–61, 2010.

[117] G. Shakhnarovich, P. A. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proceedings of the International Conference on Computer Vision*, pages 750–759, Nice, France, October 2003.

[118] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotic Research*, 22(6):371–392, 2003.

[119] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–76, Madison, WI, USA, June 2003.

[120] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):814–827, 2003.

[121] D. J. Spiegelhalter, R. Franklin, and K. Bull. Assessment, criticism, and improvement of imprecise probabilities for a medical expert system. In *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, 1989.

[122] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, volume 1, pages 605–612, Madison, WI ,USA, June 2003.

[123] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Viewpoint invariant exemplar-based 3D human tracking. *Computer Vision and Image Understanding*, 77(1-3):291–330, 2008.

[124] A. Sundaresan and R. Chellappa. Model driven segmentation of articulating humans in Laplacian Eigenspace. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1771–1785, 2008.

[125] A. Sundaresan, R. Chellappa, and R. RoyChowdhury. Multiple view tracking of humans modelled by kinematic chains. In *Proceedings of the IEEE Conference on Image Processing*, volume 2, pages 1009–1012, Singapore, October 2004.

[126] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(3):349–363, 2000.

[127] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 209(5500):2319 – 2323, 2000.

[128] N. D. Thang, T.-S. Kim, Y.-K. Lee, and S.-Y. Lee. Estimation of 3-D human body posture via co-registration of 3-D human model and sequential stereo information. *Applied Intelligence*, DOI:10.1016/j.ins.2010.02.003, 2010.

[129] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, 48(1):9–19, 2002.

[130] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1483–1489, 2008.

[131] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(11):1473–1488, 2008.

[132] M. Z. Uddin, J. J. Lee, and T.-S. Kim. Independent shape component-based human activity recognition via hidden Markov model. *Applied Intelligence*, 33(2):193–206, 2010.

[133] M. Z. Uddin, P. T. H. Truc, J. J. Lee, and T.-S. Kim. Human activity recognition using independent component features from depth images. In *Proceedings of the 5th International Conference on Ubiquitous Healthcare*, pages 181–183, Busan, Korea, November 2008.

[134] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *Proceedings of the International Conference on Computer Vision*, pages 403–410, Beijing, China, October 2005.

[135] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[136] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, HI, USA, December 2001.

[137] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Computer Graphics*, 27(3):1–9, 2008.

[138] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1–2:1–305, 2008.

[139] F. Wang and C. Zhang. Estimating anthropometry and pose from a single image. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, Minnesota, USA, June 2007.

[140] L. Wang, T. Tan, H. Ninh, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003.

[141] P. Wang and J. M. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 790–797, New York, NY, USA, June 2006.

[142] R. Wang and W. K. Leow. Human body posture refinement by nonparametric belief propagation. In *Proceedings of the IEEE Conference on Image Processing*, volume 3, pages 1272–1275, Genoa, Italy, September 2005.

[143] S. Waterhouse, D. MacKay, and T. Robinson. Bayesian methods for mixtures of experts. In *Proceedings of Advances in Neural Information Processing Systems*, pages 351–357, Denver, CO, USA, November 1995.

[144] M. Wax and T. Kailath. Detection of signals by information-theoric criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33:387–392, 1985.

[145] X. K. Wei and J. Chai. Modeling 3D human poses from uncalibrated monocular images. In *Proceedings of the International Conference on Computer Vision*, pages 1873 – 1880, Texas A&M University, USA, September 2009.

[146] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 379–385, Champaign, IL, USA, June 1992.

[147] H. D. Yang and S. W. Lee. Reconstruction of 3D human body pose from stereo image sequences based on top-down learning. *Journal of Pattern Recognition*, 40(11):3120–3131, 2007.

[148] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[149] P. Zarchan and H. Musoff. *Fundamentals of Kalman Filtering: A Pracitcal Approach*. AIAA, 2005.