Depth Map and 3D Imaging Applications:

Algorithms and Technologies

Aamir Saeed Malik
Universiti Teknologi Petronas, Malaysia

Tae-Sun Choi
Gwangju Institute of Science and Technology, Korea

Humaira Nisar
Universiti Tunku Abdul Rahman, Perak, Malaysia

Information Science
REFERENCE

Managing Director: Lindsay Johnston
Book Production Manager: Sean Woznicki
Development Manager: Joel Gamon
Development Editor: Michael Killian
Acquisitions Editor: Erika Carter
Typesetters: Mackenzie Snader
Print Coordinator: Jamie Snavely
Cover Design: Nick Newcomer

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Depth map and 3D imaging applications: algorithms and technologies / Aamir Saeed Malik, Tae Sun Choi, and Humaira Nisar, editors.
p. cm.
Summary: "This book present various 3D algorithms developed in the recent years to investigate the application of 3D methods in various domains, including 3D imaging algorithms, 3D shape recovery, stereoscopic vision and autostereoscopic vision, 3D vision for robotic applications, and 3D imaging applications"-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-61350-326-3 (hardcover) -- ISBN 978-1-61350-327-0 (ebook) -- ISBN 978-1-61350-328-7 (print & perpetual access) 1. Algorithms. 2. Three-dimensional imaging. I. Malik, Aamir Saeed, 1969- II. Choi, Tae Sun, 1952- III. Nisar, Humaira, 1970- IV. Title: Depth map and three-D imaging applications.
QA9.58.D47 2012
621.36'7015181--dc23

2011031955

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 28
Recovering 3D Human Body Postures from Depth Maps and Its Application in Human Activity Recognition

Nguyen Duc Thang
Kyung Hee University, Korea

Md. Zia Uddin
Kyung Hee University, Korea

Young-Koo Lee
Kyung Hee University, Korea

Sungyoung Lee
Kyung Hee University, Korea

Tae-Seong Kim
Kyung Hee University, Korea

ABSTRACT

We present an approach of how to recover 3D human body postures from depth maps captured by a stereo camera and an application of this approach to recognize human activities with the joint angles derived from the recovered body postures. With a pair of images captured with a stereo camera, first a depth map is computed to get the 3D information (i.e., 3D data) of a human subject. Separately the human body is modeled in 3D with a set of connected ellipsoids and their joints: the joint is parameterized with the kinematic angles. Then the 3D body model and 3D data are co-registered with our devised algorithm that works in two steps: the first step assigns the labels of body parts to each point of the 3D data; the second step computes the kinematic angles to fit the 3D human model to the labeled 3D data. The co-registration algorithm is iterated until it converges to a stable 3D body model that matches the 3D human posture reflected in the 3D data. We present our demonstrative results of recovering body postures in full 3D from continuous video frames of various activities with an error of about 6^0-14^0 in the estimated kinematic angles. Our technique requires neither markers attached to the human subject nor multiple cameras: it only requires a single stereo camera. As an application of our body posture recovery technique in 3D, we present how various human activities can be recognized with the body joint angles derived from the recovered body postures. 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. Our experimental results show the presented techniques outperform the conventional human activity recognition techniques.

DOI: 10.4018/978-1-61350-326-3.ch028

INTRODUCTION

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 map 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 3D W1 and a stereo webcam, Minoru 3D. Lately, 3D 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 3D movies and 3D TVs with the special viewing glasses is becoming a part of our lives these days.

Another area where the depth information could be valuable is the field of human computer interaction (HCI). In this area, 3D motion information of a user is utilized to better control external devices such as computers and games. In the conventional ways, capturing 3D human motion or movement (i.e., a sequence of human postures) is typically done using optical markers or motion sensors. Such systems are capable of producing some kinematic parameters of human motion with high accuracy and speed using wearable optical markers or sensors. However, it is inconvenient to a user who needs to wear specially designed optical markers or sensor-suits when running these systems. This disadvantage combined with the high cost equipment makes the systems impractical in daily use applications. In the case of using motion sensors, a user has to hand-hold controllers equipped with accelerometers or gyroscopes. One good example is the Wii controller of Nintendo which uses optical sensors and accelerometers to recognize the hand motion of the user to control the games. Lately, some efforts are being made to capture the whole body movement without the markers or motion sensors. Using a stereo camera and its derived depth map is one of options, since depth maps may provide sufficient 3D information to derive human body motions in 3D. Although this approach should open a new possibility for various novel applications in HCI such as games and u-lifecare, obtaining human body postures in 3D directly from depth maps is not very straightforward.

There have been some attempts to develop marker-less systems to estimate human motion from a sequence of monocular images or RGB images, only reflecting 2-D information. Because the 3D information of the subject is lost, the efforts to reconstruct the 3D motion of the subject from only monocular images face difficulties with ambiguity and occlusion that lead to inaccurate results (Yang & Lee, 2007). Therefore, most marker-less systems use multiple cameras to capture 3D human motion. Through such systems, the 3D information of the observed human subject is captured from different directional views, thereby providing better results of recovered human motion in 3D (Knossow et al., 2008; Gupata et al., 2008). However, it is usually complicated to setup such a system, because it requires enough space where the cameras can be installed. Also it requires synchronization of the cameras. Thus, there are always some tradeoffs between the flexibility of using a single camera and the ability to get the 3D information using multiple cameras.

Another way of recovering a series of human postures or motion in full 3D is to utilize the information in depth maps. However, there has been little effort to recover 3D human body postures using this approach. Some conventional works to estimate human body postures from depth maps

can be classified in the following two approaches: namely the *matching-based* approach and the *model-based* approach. In the *matching-based* approach (Yang & Lee, 2007), one tries to match a depth map with a set of generated human body postures to find the most compatible human body posture in the depth map. In the *model-based* approach (Urtasun et al., 2006), one creates a human body model and fits the model to the given depth map to estimate its corresponding human body posture. In this chapter, we present an approach of recovering human body postures from depth maps based on the framework of the *model-based* approach. However, in our approach we have added a novel step of detecting human body parts and incorporated it into our co-registration algorithm such that human body postures can be estimated in a more efficient and generalized framework.

The chapter begins with a survey of the conventional approaches including the use of optical markers and multiple cameras to capture 3D human body postures. We discuss their advantages and disadvantages in comparison with our approach of recovering 3D human body postures directly from depth maps without using optical markers or multiple cameras. In the following sections, we present technical details of our method with examples and demonstrations. Subsequently, as an application of our technique in human activity recognition (HAR), we present a section of how various human activities can be recognized with the derived body joint angles from the recovered body postures. We conclude the chapter with future research directions.

BACKGROUND

In general, there are two main frames of human motion (or a time-series of postures) capture systems. One is the optical system (i.e., video sensor based), which uses video cameras to obtain images and applies image processing techniques to reconstruct human motion from the acquired images. The other is the non-optical (i.e., motion sensor based) system, which uses gyroscopes (to measure angular velocity), accelerometers (to measure acceleration), or magnetic sensors (to measure the position and orientation of magnetic markers) to capture human motion. Here, we mainly focus on the systems using optical devices.

Most 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 method, the kinematic parameters are estimated using the relative locations of the detected markers. For instance, the kinematic angles at the knee joint are estimated based on the 3D 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 postures via VICON exhibits a recording frame rate up to 240 frames-per-second that is enough to capture human activities with fast movements. However the devices for this approach are very expensive.

Nowadays, there are increasing research efforts to develop a marker-less system to recover human body postures in 3D from video. Obviously, the video is conveniently recorded with a normal camera to provide a sequence of monocular images. The articulated human body model was reconstructed from some detected regions of the human body in monocular images using the inverse kinematics (Taylor, 2000). In other approaches, a probabilistic model was designed to establish the relationship between the human postures and the cues from images like color, contours, and silhouettes. Machine learning techniques such as the sampling by the Monte-Carlo method (Lee & Cohen, 2006) were applied to find the human

body posture most probabilistically compatible with the information given in the images. However, as the depth information is lost (i.e., the 3D object is projected into a 2-D image), there will be an ambiguity of reconstructing a 3D human posture from a monocular image. The appearance of a human subject in an image might also correspond to many possible configurations of the human posture in 3D. Due to this limitation, most previous researches based on a monocular image concentrate only on detecting the human body parts (Hua et al., 2005; Ramanan et al., 2007; Roberts et al., 2007).

Rather than processing on a single image, a lot of attempts have been proposed to utilize monocular images acquired with multiple cameras to get more accurate results of recovering human body postures. For instance, a setup with multiple cameras described in (Horaud et al., 2009; Knossow et al., 2007) was composed of six cameras installed at different locations to estimate motion of a tracked subject. Typically, the information in monocular images with different directional views is combined to reconstruct the 3D data of a human subject. The 3D data might be presented by 3D voxels or by a cloud of 3D points. Thus, with each presentation of 3D data, there are different ways to reconstruct human body postures. In (Sundaresan & Chellapa, 2008), the authors presented a method to segment the 3D voxels into different body parts and registered each part by one quadric surface to reconstruct the articulated human model. To segment the 3D 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 3D data. Based on this skeleton structure, they could assign the 3D data to corresponding human body parts using probabilistic registration. Some other methods like ISOMAP (Chu et al., 2003; Tenenbaum et al., 2000), Locally Linear Embedding (Roweis & Saul, 2000), or Multidimensional Scaling (Cox & Cox, 2001) are also available to recover the human skeleton structure of the 3D voxels. Meanwhile, with another form of representation of 3D data, a cloud of 3D points, in (Plankers & Fua 2003), 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 3D points, consequently finding the human body posture most fitted with the 3D data. Rather than using a single surface like the *soft object*, in (Horaud et al., 2009), they used a set of surfaces with ellipsoids to present the human body. In order to perform the registration of the ellipsoids to the 3D data, each 3D point was cast into one ellipsoid using the datum distance and the least-square estimator was utilized to draw the ellipsoids close the 3D data.

Although the marker-less systems using multiple cameras to recover human body postures can overcome the disadvantages of the system using a single camera with the ambiguities and occlusions of the 3D data when presented in a monocular image, there are still some remaining limitations in the multiple camera-based approaches. For instance, there is a need for extra software and hardware to support the transfer of large video data from multiple cameras over a network. Also, the data acquired with more than one camera must be calibrated to compute the 3D coordinate of each pixel of the recorded images within the same coordinate system. Moreover, the multiple cameras require a complicated installation. Therefore, using a single stereo camera should be more flexible and practical in the recovery of human body postures. As mentioned, there are two types of approaches of recovering human body postures from depth maps.

The first is the *matching-based* approach in which a set of human body postures is generated and compared with a depth map derived from a stereo camera to find the best matching posture. In (Yang & Lee, 2007), about 100,000 human postures, presenting most appearances of the hu-

man body in 3D, were created and stored in an exemplar database. However, with a large number of human body postures, the authors had to develop an efficient algorithm to organize and retrieve the human body posture stored in the database. To avoid generating all possible human postures, in (Olivier et al., 2009), only a limited number of human postures at the time index *t* that are close to the human body posture estimated at the time index *t-1* were generated. This method evaluated the discrepancies between the created human postures and the 3D information of the new depth map given at the time index *t* to find the human posture best compatible with the depth map. The drawback of this method is that with the limited number of generated postures, the accuracy of estimating human body postures tends to be low. In the opposite case, with the increased number of generated postures, the time needed to search for an appropriate human posture gets prolonged.

Apart from the *matching-based* approach, the *model-based* approach (Urtasun et al., 2006) estimates human body postures directly from depth maps without using a set of temporary postures for matching. This approach models an articulated human body in 3D and formulates an estimation problem to minimize the difference between the human model and the information in a depth map to recover a human posture. Our technique of recovering human body postures presented in this chapter is based on the framework of this *model-based* approach. However, we have extended and generalized the approach by developing a co-registration algorithm with an additional step of detecting human body parts in 3D before fitting the human body model to 3D data (Thang et al., 2010a).

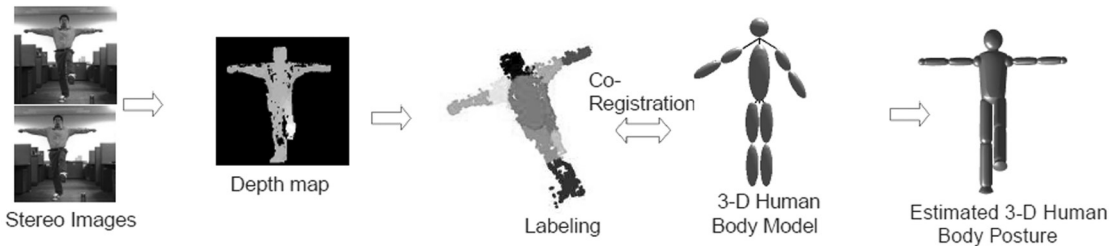
HOW TO RECOVER 3D HUMAN BODY POSTURES FROM DEPTH MAPS

The overall steps of our method of recovering human body postures from depth maps are presented in Figure 1. First, we preprocess a pair of stereo images to obtain a depth map and calculate the 3D information (i.e., 3D data) from the depth map. Separately, we create our articulated human body model using a set of ellipsoids and parameterize the model with kinematic joint angles. Finally, we co-register the body model to the 3D data of the depth map to estimate the joint angles. Our co-registration involves the following two main steps:

- **Labeling:** The labeling step assigns a label of each human body part (i.e., an ellipsoid) to each point of the 3D data using the information and cues from RGB images.
- **Model Fitting:** after the body part labeling, the model fitting step fits each point to its corresponding ellipsoid by minimizing the distance between them.

This two-step co-registration process is iterated to minimize the differences between the 3D human body model and the observed 3D data. Finally, the algorithm finds the best human pos-

Figure 1. Essential steps of our methodology of recovering 3D human body postures from depth maps



ture on a frame-by-frame basis. In the following sub-sections, more details of each process are presented.

Preprocessing of Stereo Images

As mentioned, a stereo camera is used to capture a pair of images in a time sequence containing human motion. For each pair of images, we apply the stereo matching algorithm (Cech & Sara, 2007) to compute the pixel disparities between them, generating a depth map that decodes the 3D information of the scene: the pixel with higher disparity value is closer to the camera than other pixels. Continuously, we perform the background modeling and subtraction (Wang et al., 2003) in a RGB image to get the binary silhouette of a human subject and use the binary silhouette to extract the region of interest in the depth map containing only the 3D information of the human subject. Then, for each pixel belonging to the human body region in the depth map, we calculate its coordinate in the 3D space in order to estimate the kinematic joint angles of the human posture correctly. The depth value Z_w of a pixel in the 3D coordinate system is computed by

$$Z_w = \frac{f_c b}{d} \tag{1}$$

where f_c is the focus length, b the base-line, and d a disparity value of the pixel. The two remained coordinate X_w and Y_w are computed by

$$X_w = \frac{uZ_w}{f_c}, \tag{2}$$

$$Y_w = \frac{vZ_w}{f_c}, \tag{3}$$

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

3D Human Body Modeling

We create the articulated human body with a set of ellipsoids where each ellipsoid represents one human body part as shown in Figure 2(a). For the convenience of transformation computations, we formulate the equation of each ellipsoid in the 4-D projective space as,

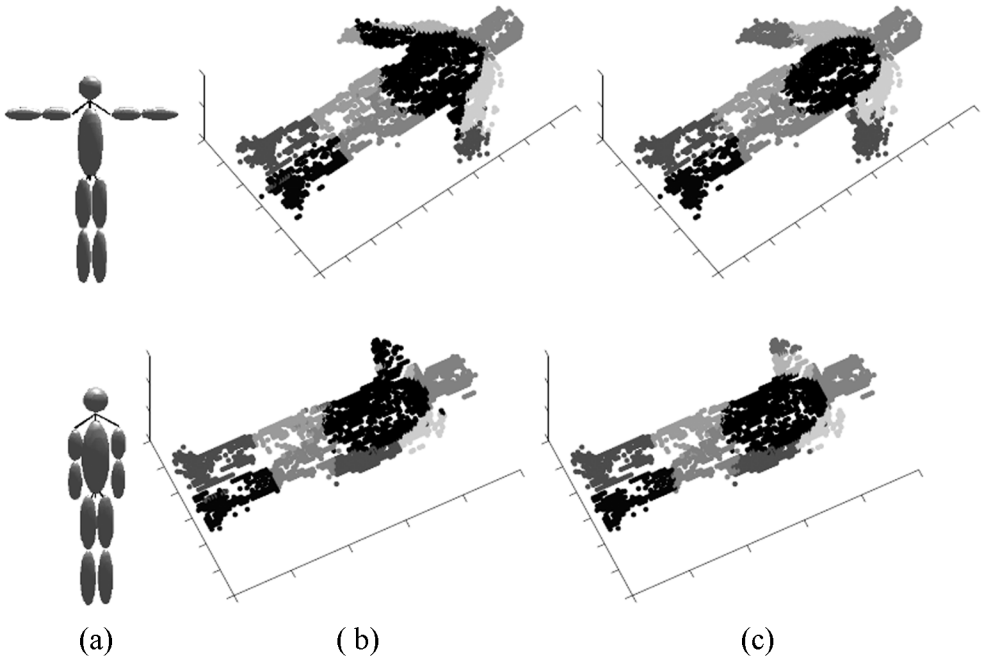
$$q(X) = X^T \mathrm{Q}_\theta^T \mathrm{S}^T \mathrm{D} \mathrm{S} \mathrm{Q}_\theta X - 2 = 0 \tag{4}$$

where the constant matrix $\mathrm{D} = diag[a^{-2}, a^{-2}, b^{-2}, 1]$, $b \geq a$ determines the size of the ellipsoid. The constant matrix $\mathbf{S}$ locates the center of the ellipsoid in the local coordinate attached to the ellipsoid. Q_θ is the skeleton-induced transformation matrix. $X = [x, y, z, 1]^T$, indicating the coordinate of a 3D point in the 4-D projective space. Each segment of the human body model is controlled by a series of transformations specified by the kinematic parameters at each body joint, therefore Q_θ is a matrix function of $\theta = (\theta_1, \theta_2, ..., \theta_n)$, where $\theta_1, \theta_2, ..., \theta_n$ are n kinematic parameters. We separate Q_θ into a series of matrices where each matrix is computed based on a single parameter,

$$\mathrm{Q}_\theta = \mathrm{Q}_n(\theta_n)\mathrm{Q}_{n-1}(\theta_{n-1})...\mathrm{Q}_1(\theta_1) \tag{5}$$

where $\mathrm{Q}_1(\theta_1), \mathrm{Q}_2(\theta_2), ..., \mathrm{Q}_6(\theta_6)$ are of six degrees of freedom (DOF) (i.e., three translations and three rotations) that determine the transformation from the global coordinate system to the local coordinate system attached at the body hip. The other matrix element, $\mathrm{Q}_i(\theta_i) = \mathrm{Tr}_i \mathrm{R}(\theta_i)$ with $i > 6$ is the transformation matrix from the local coordinate system attached to the body segment i to the local coordinate system attached to the body segment $i+1$, where Tr_i is the constant translation matrix dependent on a skeleton structure and $\mathrm{R}(\theta_i)$ is the rotation matrix at each body joint around the x-, y-, or z-axis. We can assign the value of the matrix Tr_i by an identity

Figure 2. Two examples of running E-steps to detect the body part labels. (a) Initial models. (b) The label assignments found by the first iteration of E-step. (c) The label assignments found by the last iteration of E-step.



matrix if we want to add more than one DOF to a body joint.

Our defined human model is composed of 14 body segments, nine joints (i.e., two knees, two hips, two elbows, two shoulders, and one neck), and 24 DOF (i.e., two DOF at each joint and six DOF for the transformation from the global coordinate system to the local coordinate system at the body hip). In addition, another human body model using the super quadric can be created for better display of the results as in Figures 3 and 4. The formulation of the super-quadric surface without any transformation (rotation or translation) is derived as

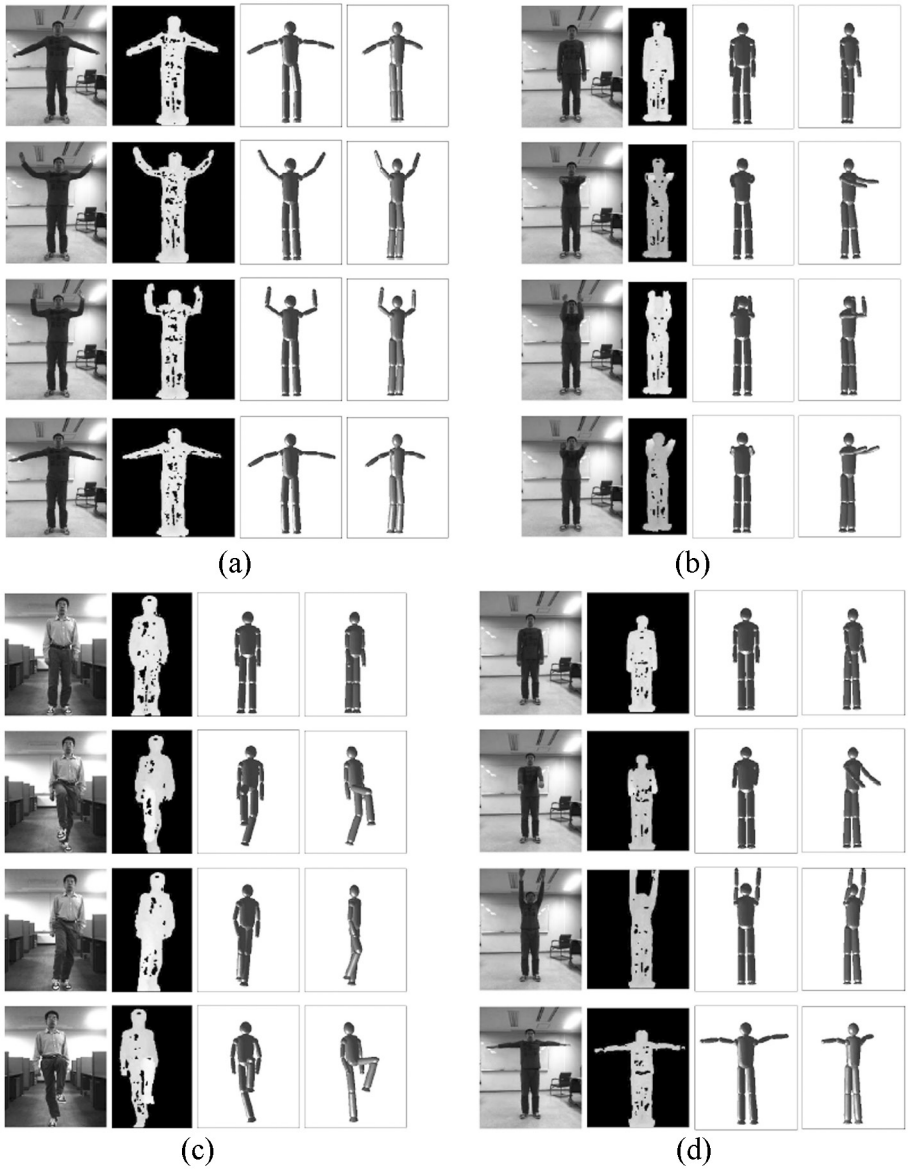
$$\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2=\left(1+\frac{sz}{c}\right)\left(1-\left(1-\frac{2z}{c}\right)^d\right) \qquad (6)$$

where a, b, and c determine the size of the super-quadric along the x-, y-, and z-axis, respectively.

Mathematical Relationship Between Human Body Model and Depth Information

In this section, we introduce a probabilistic distribution that represents the relationship between the human body posture specified by the kinematic parameters and the information in the corresponding depth map and RGB image. Let $D = (X_1,X_2,\ldots,X_N)$ denote N points of the 3D data computed from a depth map and I denote a RGB image. The supplementary variable $V = (v_1,v_2,\ldots,v_N)$ is used to label the body part where each point should belong to. The posterior probability between the label V and the kinematic parameter θ

Figure 3. Experimental results with (a) elbow movements in the horizontal direction, (b) elbow movements in the vertical direction, (c) knee movements, and (d) shoulder movements. From the left column to the right: RGB images, depth maps, and recovered human postures in the front view and $+45^0$ view.

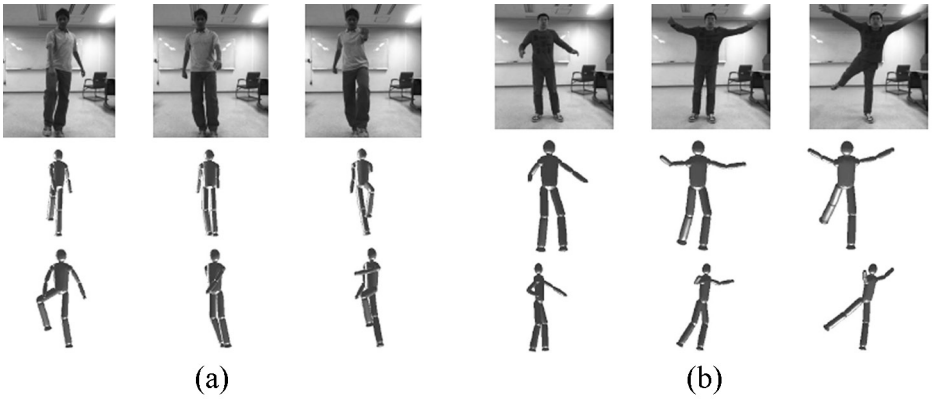


given the 3D data, D and the RGB image, I is expressed by,

$$P(V,\theta \mid I,D) \propto P(V)P(I \mid V)P(D \mid V)P(D \mid V,\theta). \quad (7)$$

Obviously, the optimal kinematic parameter θ^* that maximizes the probability distribution given in (7) represents the human body posture that is most compatible with the 3D information given in the depth map. The co-registration algorithm to estimate the optimal kinematic parameter θ^*, recovering the correct human body

Figure 4. Experimental results with (a) a walking sequence (top). Recovered human postures are depicted in the front view (middle) and -45^0 view (bottom) and (b) an arbitrary activity sequence (top). Recovered human postures are depicted in the front view (middle) and -45^0 view (bottom).



posture from the given depth map is presented in the next section.

a) Smoothness Prior

The smoothness prior found from the Potts model (Boykov et al., 2001) is given by

$$P(V) = \prod_{i=1}^{N} \prod_{j \subset N_i} P(v_i, v_j) \tag{8}$$

where N_i is a set of neighbors of the point i and $P(v_i,v_j)$ is defined by

$$P(v_i, v_j) = \begin{cases} 1 & if \ \ v_i \neq v_j \\ e^{\gamma} & if \ \ v_i = v_j \end{cases} \tag{9}$$

where γ is a positive constant. The smoothness prior $P(v_i,v_j)$ is used to derive the label of each point toward the same label of its neighbors that makes the labeling outcomes smooth and removes outliers. Here, the neighbors of one pixel in a depth map lie inside a circle with the center at the pixel's location with its radius $d=3$.

b) Image Likelihood

The RGB image containing the information of a human subject in a color space can be used to detect some human body parts, providing extra information for assigning the labels of 3D data. The detection results are integrated into equation (7) by the likelihood term $P(I|V)$,

$$P(I \mid V) = \prod_{i=1}^{N} \varphi(I \mid v_i). \tag{10}$$

In our work, we perform the face and torso detection to calculate the probability of one point inside the detected regions getting a label *'head'* or *'torso'*. The face areas are located by detecting the skin color in the HSV color space (Conaire et al., 2007).

$$\varphi(I \mid v_i = head) = \begin{cases} e^{c} & \text{the pixel } i \text{ is marked as } \textit{'face'} \\ 1 & \text{other wise} \end{cases} \tag{11}$$

where c is a positive constant.

The likelihood of a pixel labeled as *'torso'* is computed based on the function $f(r_i)$,

$$f(r_i) = \kappa e^{-d(r_i)} \tag{12}$$

where $d(r_i)$ is the algebraic distance from a point r_i in a RGB image with a coordinate $[x^r,y^r,1]^T$ in the 3D prospective space to the center of the body O_{body}: O_{body} lies in a middle between the center of the face and the center of a binary silhouette. K is a positive constant. The algebraic distance $d(r_i)$ is computed by,

$$d(r_i) = r_i^T \mathrm{Q}_e^T \mathrm{D}_e \mathrm{Q}_e r_i - 1 \tag{13}$$

where $\mathbf{D}_e$ and $\mathbf{Q}_e$ are the 3×3 matrices that configure the size and shape of the ellipse representing the torso. The likelihood to assign a point as a *'torso'* is given by,

$$\varphi(I \mid v_i = torso) = \begin{cases} f(r_i) & d(r_i) \le 1 \\ 1 & \text{other wise.} \end{cases} \tag{14}$$

c) Pairwise Geodesic Relationship among 3D Points

The geodesic distance is measured by the length of the shortest path between two points on a curved surface. During the movement and deformation of a non-rigid object like the human body, the geodesic distance between any two points on the boundary surface of the object is preserved. Therefore, we utilize this property of the geodesic distance to derive the geodesic constraints between any two points of the 3D data representing the human body.

Since there are a large number of 3D points, we need a large number of computations to estimate the geodesic distance among all pairs of the 3D points. In order to reduce the number of computations, we assign a set of close points into a group, called a *cell*. All 3D points belonging to the same cell receive the same geodesic constraint. Computing the geodesic distance by the shortest path distance in graph using the Dijkstra's algorithm (Dijkstra, 1959), we express $P(D|V)$ by

$$P(D \mid V) = \prod_{i=1}^{N} \prod_{j_c=1}^{N_c} P_{geo}(D \mid v_i, v_{j_c}) \tag{15}$$

$$P_{geo}(D \mid v_i, v_{j_c}) = \begin{cases} e^{-\alpha} & d(v_{i_c}, v_{j_c}) < d_{\min}(v_{i_c}, v_{j_c}) \\ e^{-\beta} & d(v_{i_c}, v_{j_c}) > d_{\max}(v_{i_c}, v_{j_c}) \end{cases} \tag{16}$$

where i_c is the cell that holds the point i, $d(v_{i_c}, v_{j_c})$ the geodesic distance between the cell i_c and j_c, N_c the number of cells, and (α,β) 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 bound for the geodesic distance between a pair of labels. The two related labels assigned to two 3D points that are too far or too close are penalized to reduce the belief in these assignments.

d) Reconstruction Error

The discrepancies between the human model created by a set of connected ellipsoids and the cloud of 3D points are measured by the total Euclidean distances from each 3D point to the ellipsoid corresponding to the label of this point. Thus, the Euclidean distance is considered as another factor to assign the label of each point during the registration process. $P(D \mid V, \theta)$ is defined by

$$P(D \mid V, \theta) = \prod_{i=1}^{N} e^{-\frac{d^2(X_i, \theta, v_i)}{2\sigma^2}} \tag{17}$$

where $d(X_i, \theta, v_i)$ is the Euclidean distance from the point X_i to the ellipsoid v_i and the constant σ is variance. The Euclidean distance is calculated by the distance from one point to the nearest point lying on the ellipsoid surface. In general, to compute the Euclidean distance $d(X_i, \theta, v_i)$, we need to solve a sixth-degree polynomial equation (Heckbert, 1994). However, with the symmetric ellipsoid defined in our articulated human model,

a sixth-degree polynomial equation is simplified to a fourth-degree polynomial that has an analytical solution allowing us to compute its roots.

Co-Registration of 3D Human Body Model and 3D Depth Information

A human body posture that best matches the observed 3D data is subject to the kinematic parameter θ^* that maximizes the posterior probability given in (7),

$$\theta^* = \arg\max_\theta \sum_V P(V, \theta \mid I, D). \quad (18)$$

To solve this optimization problem, the EM algorithm is a suitable choice with the incorporation of the latent variable, *V*. Let *Q(V)* be the probability distribution of the label *V*. Our algorithm to estimate a human body posture from a given depth map is formulated in an EM framework with the following two key steps:

- **E-step:** Assuming that the current value of the kinematic parameter θ is θ_{old}, E-step estimates the label assignments by computing the probability distribution $Q(V) = P(V \mid \theta_{old}, I, D)$ of the label given the information of the RGB image and the 3D data of the depth map.
- M-step: With the label assignment $Q_{old}(V)$ found by E-step, M-step maximizes $E_{Q_{old}(V)}[\log(P(V, \theta \mid I, D))]$ or equivalently minimizes the reconstruction error between the model and the cloud of 3D points to estimate a new optimal value of the kinematic parameter θ.

The two-step co-registration process is iterated to minimize the differences between the 3D model and the observed data and finally the correct human posture is found. More details of those two steps are presented as follows.

a) E-step: Labeling

It is intractable to calculate the exact distribution *Q(V)* of the label *V*. Therefore, we approximate the distribution *Q(V)* by using the mean field approach (Toyoda & Hasegawa, 2008). The logarithm of *Q(V)* is given by

$$\log Q(V) \propto \sum_{i=1}^{N} g_i(v_i) + \sum_{i=1}^{N} \sum_{j \subset N_i} g_{ij}(v_i, v_j) + \sum_{i=1}^{N} \sum_{j_c=1}^{Nc} h(v_i, v_{j_c}) \quad (19)$$

where $g_i(v_i)$ is the sum of the logarithms of the image likelihood in (10) and the reconstruction error in (17), $g_{ij}(v_i, v_j)$ the logarithm of the smooth prior in (8), and $h(v_i, v_{j_c})$ the logarithm of the geodesic constraints in (15),

$$h(v_i, v_{j_c}) = \log P_{geo}(D \mid v_i, v_{j_c}). \quad (20)$$

The probability of a pixel *i* having a label v_i, $q_i(v_i) = P(v_i \mid \theta, I, D)$ is iteratively updated until it approaches to a stable value by an equation

$$q_{i_{step+1}}(v_i) = \frac{1}{Z_{i_{step}}(v_i)} \exp\left\{ g_i(v_i) + \sum_{j \subset N_i} \sum_{v_j} q_{j_{step}}(v_j) g_{ij}(v_i, v_j) + \sum_{j_c=1}^{N_c} \sum_{v_{j_c}} q^{j_c}_{step}(v_{j_c}) h(v_i, v_{j_c}) \right\} \quad (21)$$

where $Z_{i_{step}}(v_i) = \sum_{v_i} q_{i_{step}}(v_i)$ is a normalization factor and $q^{j_c}_{step}(v_{j_c}) = E[q_{j_{step}}(v_{j_c})]$ an average probability of all pixels j belonging to the cell j_c. We use $\frac{1}{Z_{i_0}(v_i)} \exp\{g_i(v_i)\}$ as an initial value of $q_{i_0}(v_i)$. For simplification, we set $q^{j_c}_{step}(v_{j_c} = \varepsilon) = 1$ when the probability of the cell j_c belonging to the ellipsoid ε is largest and $q^{j_c}_{step}(v_{j_c}) = 0$ for $v_{j_c} \neq \varepsilon$. In Figure 2, we show two examples of

running E-step to detect the body part labels from the 3D data.

b) M-step: Model Fitting

After the probability distribution of the label variables is estimated from E-step, M-step computes a new value of the kinematic parameter θ as the solution of the optimization problem

$$\arg\max_{\theta} E_{Q(V)}[\log P(D \mid \theta, V)] \tag{22}$$

Here, we remove the terms in Equation (7) independent of θ. Equation (22) can be rewritten as

$$-\arg\max_{\theta} \sum_{\varepsilon=1}^{N_\varepsilon} \sum_{i=1}^{N} q_i(v_i = \varepsilon) d^2(X_i, \theta, v_i = \varepsilon) \tag{23}$$

$$\text{or } \arg\min_{\theta} \sum_{\varepsilon=1}^{N_\varepsilon} \sum_{i=1}^{N} q_i(v_i = \varepsilon) \left\| X_i - Z_i(\theta)^\varepsilon \right\|^2$$

where N_ε is the number of ellipsoids and $Z_i(\theta)^\varepsilon$ the nearest point of X_i lying on the surface of the ellipsoid ε. To reduce the number of computations, we set $q_i(v_i = \varepsilon) = 1$ for ε satisfying $q_i(v_i = \varepsilon) \geq q_i(v_i \neq \varepsilon)$ and $q_i(v_i) = 0$ for $v_i \neq \varepsilon$. We solved the non-linear optimization problem in (23) by the Levenberg-Marquardt method (Murray et al., 1994; Sundaresan et al., 2004).

To summarize, we describe the presented algorithm in Table 1.

Results of Recovering Human Body Postures from Depth Maps

In our experiments, we used a stereo camera, Bumblebee 2.0 of Point Grey Research, to capture stereo image pairs with their resolution at 640 × 480. We asked our subjects to perform various motions in front of the stereo camera as depicted in Figure 3. Note that a sequence of frames in a video stream is shown from top-to-bottom in a column. In Figures 3(a) and 3(b), the movements of the elbows in the horizontal and vertical directions were evaluated in our experiments. The subjects raised their hands up to create an angle about 90^0 between the upper hand and lower hand, then brought their hands down. In the next experiment shown in Figure 3(c), the subject in video performed an activity at their knee joints. The subject lifted his right leg up to a 90^0 between the upper leg and lower leg then he did the same motion with the other leg. In addition, we considered the body movements created by the combination of the two kinematic angles at the shoulders as in Figure 3(d). To evaluate the reconstruction error, we generated the ground-truth of the estimated kinematic angles by using the hand-label method (Gupta et al., 2008; Lee & Cohen, 2006). Some points were hand-labeled to determine the position of the body joints in the RGB images such as hand, elbow, shoulder, etc. Using the 3D information estimated from the depth maps, we computed the coordinate of these labeled points in 3D and then calculated the ground-truth angles. Then, we compared the kinematic angles of the recovered human body postures against the ground-truth angles and obtained the mean error of about 6^0~14^0 in the estimated kinematic angles.

In order to track the movements of the whole human body, the subjects were asked to perform complicated activities with all arms and legs. Figure 4 shows two video sequences and the recovered human body postures reflected in those sequences in two view angles. The average distance between the 3D points and the ellipsoids of the human model were used to evaluate the error measurements of the reconstructed postures. The average distance D_t of the frame t was computed by

$$D_t = \sum_{i=1}^{N} d_t(i) / N \tag{24}$$

Table 1. The co-registration algorithm used to estimate human body postures from depth maps

1. At the time index *t*, initialize the value of the kinematic parameter θ_t with the value of the kinematic parameter θ_{t-1} estimated at the time index *t-1*

2. **E- step:** Compute $g_i(v_i)$ from the sum of the logarithms of the image likelihood in (10) and the reconstruction error in (17) and use $\exp\{g_i(v_i)\} / Z_{i_0}(v_i)$ as an initial value of $q_{i_0}(v_i)$

3. Compute $g_{ij}(v_i, v_j)$ from the logarithm of the smooth prior in (8) and $h(v_i, v_{j_c})$ from the logarithm of the geodesic constraints in (15)

4. Update

$$q_{i_{step+1}}(v_i) = \frac{1}{Z_{i_{step}}(v_i)} \exp\left\{ g_i(v_i) + \sum_{j \subset N_i} \sum_{v_j} q_{j_{step}}(v_j) g_{ij}(v_i, v_j) + \sum_{j_c=1}^{N_C} \sum_{v_{j_c}} q_{step}^{j_c}(v_{j_c}) h(v_i, v_{j_c}) \right\}$$

5. If $q_{i_{step+1}}(v_i)$ has not converged, go back to step 3

6. **M-step:** Estimate new values of the kinematic parameter

$$\theta_t = \arg\min_\theta \sum_{\varepsilon=1}^{N_\varepsilon} \sum_{i=1}^{N} q_i(v_i = \varepsilon) \left\| X_i - Z_i(\theta)^\varepsilon \right\|^2$$

7. If θ_t has not converged, go back to step 2

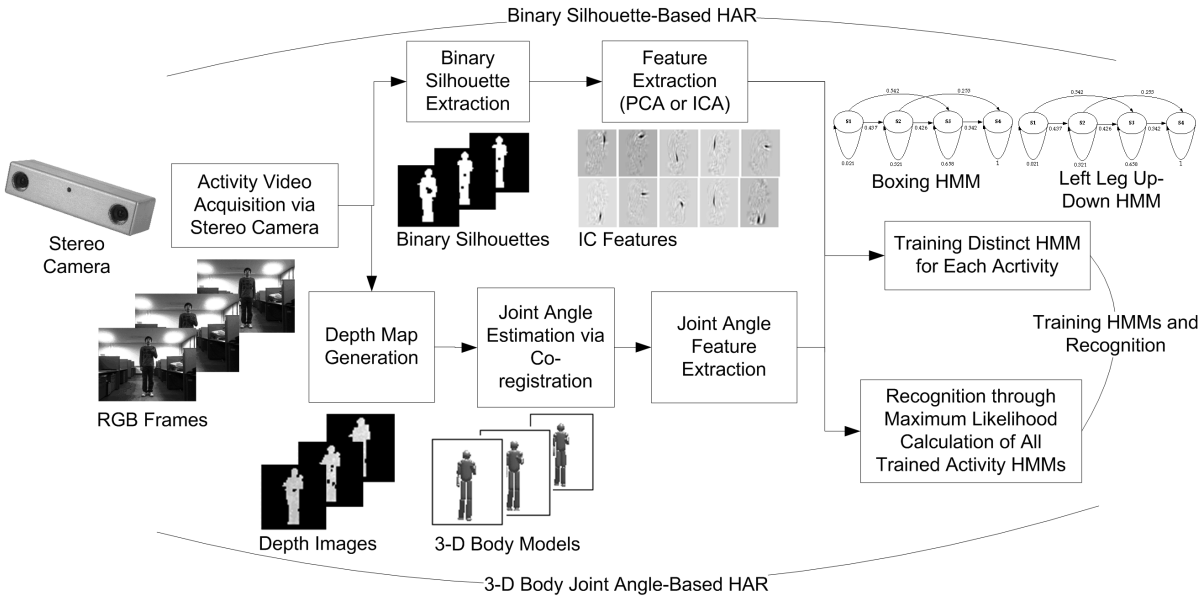
where $d_t(i)$ is the Euclidean distance between the point *i* and the nearest ellipsoid of the human model and *N* is the number of points. The mean error distance D_t for the walking and arbitrary sequences depicted in Figure 4 came out to be 0.06m and 0.04m respectively.

HUMAN ACTIVITY RECOGNITION USING BODY JOINT ANGLES

Human Activity Recognition (HAR) is defined as recognizing various human activities utilizing external sensors such as acceleration, motion, or video sensors. In recent years, HAR from video has evoked considerable interests among researchers in computer vision and image processing communities (Robertson & Reid, 2006). A key reason for this is its potential usefulness of the outcomes of such recognition in practical applications such as human computer interaction, automated surveillance, smart home, and human healthcare applications. A general method for video-based HAR starts with the extraction of key 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.

In general, 2-D binary silhouettes of human body shapes are the most common representations of human activity that have been applied for video-based HAR (Yamato et al., 1992; Carlsson & Sullivan, 2002; Niu & Abdel-Mottaleb, 2004; Niu & Abdel-Mottaleb, 2005; Uddin et al., 2008; Uddin et al., 2009). For instance, in (Yamato et al., 1992), 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 Hidden Markov Models (HMMs). In (Carlsson & Sullivan, 2002), a silhouette matching key frame-based approach was applied to recognize forehand and backhand strokes from tennis videos. Regarding binary silhouette-based features, Principal Com-

Figure 5. Processes involved in the binary silhouette and 3D body joint angle-based HAR



ponent Analysis (PCA), a feature extractor based on the second-order statistics, is most commonly applied (Niu & Abdel-Mottaleb, 2004; Niu & Abdel-Mottaleb, 2005; Uddin et al., 2009). 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 (Niu & Abdel-Mottaleb, 2004; Niu & Abdel-Mottaleb, 2005), the authors utilized PC features from binary silhouettes and optical flow-based motion features in combination with HMM to recognize different view-invariant activities. The top flow of Figure 5 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, are then applied to a recognition technique 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 postures 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 postures in different activities. In this regard, depth silhouettes for human body representations can be a solution. In the case of depth-based silhouette representation, the pixel values are set on the basis of the distance to the camera and hence it can provide better activity information than the binary silhouettes. In (Uddin et al., 2008; Uddin et al., 2009), the authors proposed to use a new feature extraction technique called Independent Component Analysis (ICA) to produce prominent local features from time-sequential depth silhouettes to be used with HMMs and obtained superior HAR performance than the binary silhouette-based approaches.

However, depth silhouettes do not convey truly 3D information of the human body postures and hence generates the similar problems as binary silhouettes: they represent the human body in dif-

ferent activities from one angle view of depth. As the human body consists of limbs connected with joints, if one is able to obtain their 3D joint angle information, one can form much stronger features than conventional silhouette features that will lead to significantly improved HAR. In this section, we present an application of HAR based on our estimated 3D body joint angle features and HMM. From the time-sequential activity video frames, the joint angles are first estimated by co-registering a 3D 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 Figure 5 shows the basic processes regarding 3D 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. Some more details of the essential processing steps are given below.

3D Joint Angle Features in Human Activities

Once we obtain the joint angles of the 3D human body for each video frame as discussed earlier, we can utilize these to represent various human activities effectively. The estimated joint angles from a video frame of a particular activity form a feature vector: thus, each activity video clip is represented in a sequence of joint angle feature vectors as $(F_1,F_2,...,F_T)$, where T is the length of the activity video. Therefore, the 3D 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.

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 (Lawrence & Rabiner, 1989; Niu & Abdel-Mottaleb, 2004; Niu & Abdel-Mottaleb, 2005; Uddin et al., 2008; Uddin et al., 2009). Basically, 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. To learn a video-based human activity in a 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 (Lawrence & Rabiner, 1989).

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

$$decision = \arg\max_{i=1,2,\ldots,M}\{P(O \mid H_i)\} \quad (25)$$

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 (Uddin et al., 2008; Uddin et al., 2009).

Results of Recognizing Various Human Activities

We had built a database of six different activities (namely, left hand up-down, right hand up-down, both hands up-down, boxing, left leg up-down, and right leg up-down) to be trained and recognized via our 3D 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.

We started our experiments with the traditional binary silhouette-based HAR. Table 2 shows the experimental results of HMM-based HAR utilizing the IC features of binary silhouettes and joint angle features of 3D body model respectively. As ICA is superior to PCA by extracting the local binary silhouette features (Uddin et al., 2009), it was utilized for HAR where 150 features were considered in the feature space. Binary silhouettes were not appropriate to recognize the activities used in our experiments, yielding a much lower mean recognition rate of 58.33%. On the contrary, utilizing the 3D body joint angle features, we obtained a mean recognition rate of 92.50%, which is far better than that of the binary silhouette-based HAR. The experimental results show that the 3D 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.

FUTURE RESEARCH DIRECTIONS

As presented, our human motion capturing system using a stereo camera is potentially applicable to various biomedical and HCI areas. However, due to the existing errors of recovered kinematic angles, our system might face difficulty with practical applications requiring high accurate results of estimating motion. For instance, in biomechanics

Table 2. Experimental results of video-based HAR using binary silhouettes vs. joint angles

Approach	Activity	Recognition Rate	Mean	Standard Deviation
Binary Silhouette-Based HAR	Left hand up-down	47.50%	58.33	16.78
	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		
Joint Angle-Based HAR	Left hand up-down	87.50	92.50	4.18
	Right hand up-down	97.5		
	Both hands up-down	87.50		
	Boxing	95		
	Left leg up-down	92.50		
	Right leg up-down	95		

measurements, some systems need small errors of recovered kinematic angles in order to analyze the detailed motion of a tracked subject. In health care areas, a human motion capturing system can be used to help a handicap person to learn how to walk, run, etc. However, the system with large errors of estimated kinematic angles might cause adverse effects to the treatment of the patient. 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 postures between two consecutive frames. A part of information used to assign the label of 3D 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 postures in a video sequence. The concerns are addressed by developing better labeling method with investigating more information to detect human body parts from RGB images as exampled in (Ninh et al., 2009). Also in the model fitting part of our algorithm, a large number of 3D points processed in the algorithm slow down the co-registration process and take into account outliers in computations that affect the recovering results. To mitigate this problem, we recently suggested a way of utilizing clusters of 3D points being assigned the same label of a body part and computing the kinematic parameters with a small number of clusters (Thang et al., 2010b). This greatly reduced the computational time, eliminated the presence of outliers, and made the presented techniques more practical.

As a practical application, we presented our work of HAR using the derived feature of joint angles, which proved its superior performance over the conventional feature of body silhouettes. We believe that our presented work in this chapter should be able to find its use in other applications such as advanced HCI, video games, smart homes, smart hospitals, etc.

CONCLUSION

In this chapter, we have presented our markerless system to recover human body postures in 3D from a sequence of depth maps acquired by a single stereo camera. We have described our methodology including how to estimate the 3D data of a depth map, how to create a human body model, and how to co-register the human body model to the 3D data. Our experimental results with real video data have shown that our method successfully recovers human body postures from depth maps: our validation indicates an error range of about 6^0-14^0 in the estimated joint angles. In addition, as an application of our technique, we have presented a HAR work using the derived body joint angles. Again our experimental results with real video data show that our HAR system produces significantly better recognition rates than the conventional approaches in which binary silhouettes are utilized to recognize human activities.

ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0003)).

REFERENCES

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1222–1239. doi:10.1109/34.969114

Carlsson, S., & Sullivan, J. (2002). Action recognition by shape matching to key frames, *IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*, (pp. 263-270).

Cech, J., & Sara, R. (2007). Efficient sampling of disparity space for fast and accurate matching, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, (pp. 1-8).

Chu, C.-W., Jenkins, O. C., & Mataric, M. J. (2003). Markerless kinematic model and motion capture from volume sequences. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, (pp. 475-482).

Conaire, C. O., O'Connor, N. E., & Smeaton, A. F. (2007). Detector adaption by maximizing agreement between independent data sources. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* (pp. 1-6).

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling*. Boca Raton, FL: Chapman and Hall.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numer Math*, *1*, 269–271. doi:10.1007/BF01386390

Gupta, A., Mittal, A., & Davis, L. S. (2008). Constraint integration for efficient multiview pose estimation with self-occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(3), 493–506. doi:10.1109/TPAMI.2007.1173

Heckbert, P. S. (1994). *Graphics germs IV*. San Diego, CA: Academic Press.

Horaud, R., Niskanen, M., Dewaele, G., & Boyer, E. (2009). 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. doi:10.1109/TPAMI.2008.108

Hua, G., Yang, M., & Wu, Y. (2005). Learning to estimate human pose with data driven belief propagation. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* (pp. 747-754).

Knossow, D., Ronfard, R., & Horaud, R. (2008). Human motion tracking with a kinematic parameterization of extremal contours. *International Journal of Computer Vision*, *79*(3), 247–269. doi:10.1007/s11263-007-0116-2

Lawrence, R., & Rabiner, A. (1989). Tutorial on hidden Markov models and selected applications in speech recognition . *Proceedings of the IEEE*, *77*(2), 257–286. doi:10.1109/5.18626

Lee, M. W., & Cohen, I. (2006). A model-based approach for estimating human 3D poses in static images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(6), 905–916. doi:10.1109/TPAMI.2006.110

Murray, R. M., Li, Z., & Sastry, S. S. (1994). *A mathematical introduction to robotic manipulation*. Boca Raton, FL: CRC Press.

Ninh, H., Han, T. X., Walther, D. B., Liu, M., & Huang, S. (2009). Hierarchical space-time model enabling efficient search for human actions. *IEEE Transactions on Circuits and Systems for Video Technology*, *19*(6), 808–820. doi:10.1109/TCSVT.2009.2017399

Niu, F., & Abdel-Mottaleb, M. (2004). View-invariant human activity recognition based on shape and motion features. *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering* (pp. 546-556).

Niu, F., & Abdel-Mottaleb, M. (2005). HMM-based segmentation and recognition of human activities from video sequences. *Proceedings of IEEE International Conference on Multimedia & Expo* (pp. 804-807).

Olivier, B., Pascal, C. C., & Arnaud, B. (2009). Fast nonparametric belief propagation for real-time stereo articulated body tracking. *Computer Vision and Image Understanding*, *113*(1), 29–47. doi:10.1016/j.cviu.2008.07.001

Plankers, R., & Fua, P. (2003). Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(9), 1182–1187. doi:10.1109/TPAMI.2003.1227995

Ramanan, D., Forsyth, D. A., & Zisserman, A. (2007). Tracking people by learning their appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(1), 65–81. doi:10.1109/TPAMI.2007.250600

Roberts, T. J., McKenna, S. J., & Ricketts, I. W. (2007). Human pose estimation using partial configurations and probabilistic regions. *International Journal of Computer Vision*, *73*(3), 285–306. doi:10.1007/s11263-006-9781-9

Robertson, N., & Reid, I. (2006). A general method for human activity recognition in video . *Computer Vision and Image Understanding*, *104*(2), 232–248. doi:10.1016/j.cviu.2006.07.006

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326. doi:10.1126/science.290.5500.2323

Sundaresan, A., & Chellapa, R. RoyChowdhury, R. (2004). Multiple view tracking of humans modeled by kinematic chains. *Proceedings of IEEE International Conference on Image Processing,* (pp. 1009-1012).

Sundaresan, A., & Chellapa, R. (2008). Model driven segmentation of articulating humans in Laplacian Eigenspace. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(10), 1771–1785. doi:10.1109/TPAMI.2007.70823

Taylor, C. J. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, *80*(3), 349–363. doi:10.1006/cviu.2000.0878

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323. doi:10.1126/science.290.5500.2319

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

Thang, N. D., Kim, T.-S., Lee, Y.-K., & Lee, S. (2010b). Fast 3D human motion capturing from stereo data using Gaussian clusters. *International Conference on Control, Automation and Systems*, (pp. 1428-1431).

Toyoda, T., & Hasegawa, O. (2008). Random field model for integration of local information and global information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(8), 1483–1489. doi:10.1109/TPAMI.2008.105

Uddin, M. Z., Lee, J. J., & Kim, T.-S. (2009). Independent shape component-based human activity recognition via hidden Markov model. *Applied Intelligence*, *33*(2). doi:.doi:10.1007/s10489-008-0159-2

Uddin, M. Z., Truc, P. T. H., Lee, J. J., & Kim, T.-S. (2008). Human activity recognition using independent component features from depth images, *Proceedings of the 5th International Conference on Ubiquitous Healthcare*, (pp. 181-183).

Urtasun, R., Fleet, D., & Fua, P. (2006). Temporal motion models for monocular and multiview 3D human body tracking. *Computer Vision and Image Understanding*, *104*(2), 157–177. doi:10.1016/j.cviu.2006.08.006

Wang, L., Tan, T., Ninh, H., & Hu, W. (2003). Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(12), 1505–1518. doi:10.1109/TPAMI.2003.1251144

Yamato, J., Ohya, J., & Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model, *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, (pp. 379-385).

Yang, H. D., & Lee, S. W. (2007). Reconstruction of 3D human body pose from stereo image sequences based on top-down learning. *Journal of Pattern Recognition*, *40*(11), 3120–3131. doi:10.1016/j.patcog.2007.01.033

ADDITIONAL READING

Besl, P., & McKay, N. (1992). A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 239–256. doi:10.1109/34.121791

Bregler, C., Malik, J., & Pullen, K. (2004). Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, *56*(3), 179–194. doi:10.1023/B:VISI.0000011203.00237.9b

Cedras, C., & Shah, M. (1995). Motion-based recognition: A survey. *Image and Vision Computing*, *13*(2), 129–155. doi:10.1016/0262-8856(95)93154-K

Chang, I., & Lin, S.-Y. (2010). 3D human motion tracking based on a progressive particle filter. *Journal of Pattern Recognition*, *43*(10), 3621–3635. doi:10.1016/j.patcog.2010.05.003

Cheung, K., Baker, S., & Kanade, T. (2005). Shape-from-silhouette across time part I: Theory and algorithm. *International Journal of Computer Vision*, *62*(3), 221–247. doi:10.1007/s11263-005-4881-5

Corazza, S., Mündermann, L., Gambaretto, E., Ferrigno, G., & Andriacchi, T. P. (2010). Markerless motion capture through visual hull, articulated ICP and subject specific model generation. *International Journal of Computer Vision*, *87*(1-2). doi:10.1007/s11263-009-0284-3

Darby, J., Li, B., & Costen, N. (2010). Tracking human pose with multiple activity models. *Journal of Pattern Recognition*, *43*(9), 3042–3058. doi:10.1016/j.patcog.2010.03.018

Dimitrijevic, M., Lepetit, V., & Fua, P. (2006). Human body pose detection using Bayesian spatio-temporal templates . *Computer Vision and Image Understanding*, *104*(2-3), 127–139. doi:10.1016/j.cviu.2006.07.007

Felzenswalb, P., & Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, *61*(1), 55–79. doi:10.1023/B:VISI.0000042934.15159.49

Forsyth, D. A., & Pronce, J. (2003). *Computer vision - a modern approach*. Upper Saddle River, New Jersey: Prentice Hall.

Fossati, A., Dimitrijevic, M., Lepetit, V., & Fua, P. (2010). From canonical poses to 3D motion capture using a single camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1165–1181. doi:10.1109/TPAMI.2009.108

Kakadiaris, I. A., & Metaxas, D. (1998). Three-dimensional human body model acquisition from multiple views. *International Journal of Computer Vision*, *30*(3), 191–218. doi:10.1023/A:1008071332753

Lee, H. J., & Chen, Z. (1985). Determination of 3D human body posture from a single view. *Computer Vision Graphics and Image Processing*, *30*(2), 148–168. doi:10.1016/0734-189X(85)90094-5

Lee, M. W., & Nevatia, R. (2009). Human Pose Tracking in Monocular Sequence Using Multilevel Structured Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 27–38. doi:10.1109/TPAMI.2008.35

Martin, F., & Horaud, R. (2002). Multiple camera tracking of rigid objects. *The International Journal of Robotics Research*, *21*(2), 97–113. doi:10.1177/027836402760475324

Mikic, I., Trivedi, M. M., Hunter, E., & Cosman, P. C. (2003). Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, *53*(3), 199–223. doi:10.1023/A:1023012723347

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2), 90–126. doi:10.1016/j.cviu.2006.08.002

Mori, G., & Malik, J. (2006). Recovering 3D human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(7), 1052–1062. doi:10.1109/TPAMI.2006.149

O'Rourke, J., & Badler, N. I. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*(6), 522–536.

Ong, E.-J., Micilotta, A. S., Bowden, R., & Hilton, A. (2006). Viewpoint invariant exemplar-based 3D human tracking. *Computer Vision and Image Understanding*, *104*(2), 178–189. doi:10.1016/j.cviu.2006.08.004

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, *108*(1-2), 4–18. doi:10.1016/j.cviu.2006.10.016

Roberts, T. J., McKenna, S. J., & Ricketts, I. W. (2007). Human pose estimation using partial configurations and probabilistic regions. *International Journal of Computer Vision*, *73*(3), 285–306. doi:10.1007/s11263-006-9781-9

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1-3), 7–42. doi:10.1023/A:1014573219977

Shankar, R. R., Allen, Y. Y., Shankar, S., & Yi, M. (2010). Robust algebraic segmentation of mixed rigid-body and planar motions from two views. *International Journal of Computer Vision*, *88*(3), 425–446. doi:10.1007/s11263-009-0314-1

Sidenbladh, H., & Black, M. J. (2003). Learning the statistics of people in images and video. *International Journal of Computer Vision*, *54*(1-3), 181–207. doi:10.1023/A:1023765619733

Signal, R., Balan, A., & Black, M. J. (2010). HumanEva: Synchronised video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, *87*(1-2), 4–27. doi:10.1007/s11263-009-0273-6

Sminchisescu, C., Kanaujia, A., & Metaxas, D. (2006). Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, *104*(2-3), 210–220. doi:10.1016/j.cviu.2006.07.014

KEY TERMS AND DEFINITIONS

Depth Map: A 2-D image representing the depth information of a scene using gray-scaled colors.

Human Computer Interaction (HCI): A research of interaction between users and computers.

Maker-Based Human Motion Capture: An approach of capturing human motion by attaching markers to the human body. The trajectories of the markers detected in 3D space provide the motion information of the tracked subject.

Markerless-Based Human Motion Capture: An approach of capturing human motion without using markers.

Stereo Camera: A type of camera composed of two or more lenses to allow taking some pictures of a scene in alternate view angles to estimate the information of depth.

Stereo Matching Algorithm: An algorithm used to generate the depth map from a pair of images captured by a stereo camera.

Stereopsis: A process of combining two images received from two human eyes to create a 3D sensation about viewed objects.