Proceeding of International Conference on

uHealthcare 2008

Biomedical Engineering for u-Healthcare

October 29-31, 2008
ARPINA Hotel, Busan, KOREA

Organized

Institute of Medical & Biological Engineering, SNUMRC, Seoul National University
Advanced Biometric Research Center, Seoul National University
Center for Emergency Medical Informatics, Yonsei University
Ubiquitous Biomedical System Development Center, Chungbuk National University
Ubiquitous Life Care Research Center, Kyunghee University

Technical Sponsors

Korean Society of Medical & Biological Engineering
IEEE EMBS

3-D Human Pose Estimation from 2-D Depth Images

Nguyen Duc Thang, Phan T. H. Truc, Young-Koo Lee, Sungyoung Lee, and Tae-Seong Kim

Abstract — **In this paper, we present a new method to estimate 3-D human body configuration (i.e., pose) from 2-D depth images. Our work involves modeling human body in 3-D, finding an initial body model, and estimating body configuration by matching depth information to the model. Among these steps, the most challenging problem is to find a most suitable 3-D human model among a large number of possible 3-D human pose configurations and matching the model in parts to the depth information. To solve this problem, we propose a combination approach with two stages. On the first stage, we find the approximate model correspondent to the depth image by matching the depth information between the database's model and the image. Then, we match the model coarsely to fit the body parts using an iterative Expectation Minimization (EM) algorithm. In results, we show that our method can estimate 3-D human body configuration from 2-D depth images with both simulated and real data from a depth stereo camera.**

I. Introduction

Recovering human pose from video data, or a 3-D articulated model fitting with human pose reflected in one or several images, is one of the most critical problems in computer vision and solving it brings a very basic foundation to develop much broader applications such as human activity recognition, human tracking, and human motion analysis. Many researches have been working on estimating 3-D human poses from 2-D images however they do not give us high accuracy due to the effects of occlusion when projecting 3-D objects on the camera view plane. To mitigate this problem recently, there have been some interesting works with depth information gathered from a system of multiple cameras or stereo cameras: For a multiple camera system [1] [2], we need to deploy a system of camera on different locations and a complicated calibration process to synchronize the activity of different cameras. With the same function, the stereo camera is also designed to capture the stereo view of environment: it is more flexible to deploy and robust to work on various environment conditions. The distance to each object is presented by the gray values in depth images.

Our proposed method attempts to fit the articulated human body model to the gray silhouettes in depth images rather than the binary silhouettes as the ordinary way of using the normal camera. The nearest approach can be found in [3] where the authors developed some searching schemes to match the query images with the templates or their pose configurations stored in the hierarchy database. This approach requires an optimally organized database for efficiently searching and a large storage to store the huge number of configurations. To overcome these disadvantages, a novel solution has been considered in this paper: Rather storing all possible configurations in the database, we only use limited number of initial models. To match these models with the given depth images, we derive depth images from the models and then compare them based on the *surface context*. After getting the model corresponding with the depth image, we adjust the model parts by the two-step EM algorithm to minimize the error between the model parts and their real observations. Our preliminary results show that it is feasible to estimate 3-D configuration of human body from only depth images.

N. D. Thang, P. T. H. Truc, Y. K. Lee, and S. Y. Lee are with the Department of Computer Engineering, Kyung Hee University, Rep. of Korea 446-701. T.-S. Kim is with the Department of Biomedical Engineering, Kyung Hee University, South Korea 446-701 (e-mail: tskim@khu.ac.kr).

II. Methodology

A. Overview of the Proposed Estimation System

The overall system can be described by the illustrated steps in Fig. 1, from finding the initial model to the model fitting step aiming at refining the model. In the model fitting step, we consider the depth image as a cloud of points in the 3-D space. The model fitting is an iterative procedure, including Labeling (E-Step) in which the label is assigned for each observed point, and the nearest corresponding point on ellipsoids is found (after defining the labels) with these points, and Optimization (M-Step) in which the distance between data point and its nearest point on ellipsoid model is minimized.

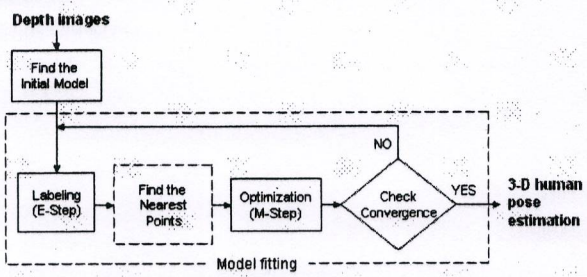


Fig. 1. Overview of our proposed system

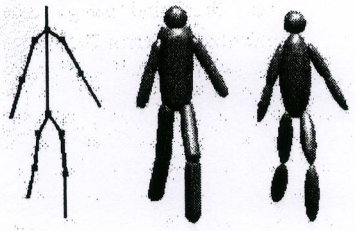
B. 3-D Human Body Model

A 3-D articulated human model is presented by a set of kinematic chains attached with ellipsoids or super-quadric surfaces as shown in Fig. 2 (b). The super-quadric surface [10] is given such as

$$\left(\frac{x}{x_0}\right)^2+\left(\frac{y}{y_0}\right)^2=\left(1+\frac{sz}{z_0}\right)\left(1-\left(1-\frac{2z}{z_0}\right)^d\right), \quad (1)$$

$0 \le z \le z_0$.

where x_0, y_0 and z_0 are the size of super-quadric along the x-axis, y-axis, and z-axis, s and d are the parameters used to control the shape of the super-quadric. In general, super-quadric models more look like human body but difficult in creation, so that ellipsoids are chosen for efficient computation and super-quadrics for making templates in the database. Each part of body has its local coordinate and is influenced by the transformation of the previous segments. We have created our model with 14 segments and 10 joints as shown in Fig. 2 (a). The root of model lies at hip. Each blob can be assigned some ellipsoids or super-quadrics, but to make simple, we only use one for each. The kinematic parameter Θ of model includes all Euler angles connected between different segments, and also free transformation from the global coordinate to the local coordinate at root. This leads to 22 Degree of Freedoms (DOFs): more particularly, 1 for each hand, 3 for each shoulder, 2 for neck, 1 for each leg, 2 for each side of hip, and 6 DOFs for one free transformation to the center of body.



(a) (b) (c)

Fig. 2. 3-D human body modeling. (a) Skeleton presenting joints and segments. (b) Super-quadric model. (c) Ellipsoid model.

C. *Finding the Initial Model from Database*

Fig. 3 presents four possible poses of the left hand of the initial human model. It is possible to make more complicated exemplars to find better approximation model at the cost of more searching time in matching. We have 4 poses for each hand, 4 poses for each leg and 3 view directions (front side, left side, right side), in total, 768 possible human pose configurations for all.

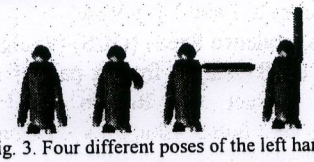


Fig. 3. Four different poses of the left hand.

The depth images are extracted from each model in the database for later comparison, so we need an efficient matching algorithm to match these depth images with the one from camera. To do this we utilize the concept of *shape context* that has been applied for image retrieval [4]. The shape context defines the relationship between each point in the shape and the rest by computing the histogram on log-polar bin (counting number of points in each bin).

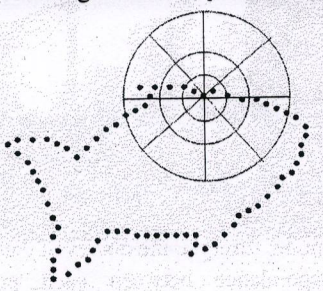


Fig. 4. Shape context with log-polar histogram defined in the shape.

We develop the idea of the shape context to the surface context to handle depth images. The process of computing the shape context is shown in Fig. 5. On the preprocessing step, we sample some points in the area bound by the human silhouette. These points combining with its gray values (depth) are connected together to create a mesh in 3-D space. The summarization of the normal vectors of sample points belonging to a bin k is

$$\hat{n}_i^k = \sum_{q_j \in Q} \hat{n}_j \quad (2)$$

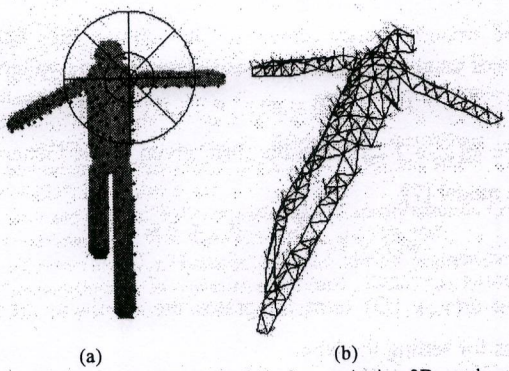
where $Q = \{q_j \neq p_i, (q_j - p_i) \in bin(k)\}$. p_i and q_j are sample points. $\hat{n}_j$ denotes the normal vector of point q_j. The surface context defines the descriptor for point p_i as

$$\hat{v}_i = \left\langle n_i^{1,x}, n_i^{1,y}, n_i^{1,z}, n_i^{2,x}, n_i^{2,y}, n_i^{2,z}, \ldots, n_i^{d,x}, n_i^{d,y}, n_i^{d,z} \right\rangle. \quad (3)$$

The distance between two points i and j in two different depth images can be computed by

$$d(p_i, p_j) = \left\| \hat{v}_i - \hat{v}_j \right\|_2. \quad (4)$$

The details of comparing two surface contexts, similar to the shape context, can be found in [5].



(a) (b)

Fig. 5. Surface context. (a) Log-polar bin of one point in a 2D area bound by the human shape. (b) Sample points connected by 3D mesh and their normal vectors.

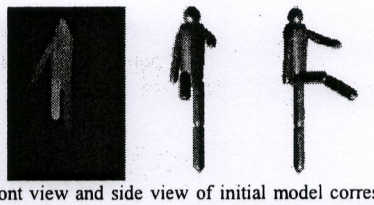


Fig. 6. The front view and side view of initial model corresponding to the depth image.

D. *Model Fitting Algorithm*

To perform the model fitting to the observed data, we need to find the correspondence between each point of the observation with one point of the model by the nearest distance. In order to increase accuracy and computation performance, each point is labeled to one part of the human body (i.e., ellipsoid). With this labeling step, we can avoid searching the corresponding point on the whole body.

The labels are calculated using the Conditional Random Fields (CRFs) model [6]. Given the observed data D, in our case, depth images, CRFs can be defined by the condition distribution of label V given D by

$$P(V|D)=\frac{1}{Z_{std}}\exp\left\{\sum_{i\in S} f_i(v_i|D)+\sum_{i\in S}\sum_{j\in N_i} f_{ij}(v_i,v_j|D)\right\}, \quad (5)$$

where f_i and f_{ij} are the data and smooth energy, N_i is a set of neighbors of i, set V includes a label of each point of D, getting the values from $X=\{X_1,X_2...,X_l\}$ representing the head, hand, torso, etc.

The data energy, expressing the relationships between the labels and depth image features, is defined as the Mean Square Error (MSE) between the observed data and the ellipsoid human model,

$$E_1(\Theta,V)=\sum_{i\in S} f_i(v_i|D)=\sum_{i\in S}{\|d_{v_i}\|^2}/{2\sigma^2}, \quad (6)$$

where Θ is the model parameters, d_{v_i} denotes the Euclidean distance between point i and the nearest point on the ellipsoid v_i, σ is the constant.

The smooth energy represents the relationships between labels of neighboring sites and is separated into two terms,

$$f_{ij}(v_i,v_j|D)=\varphi(v_i,v_j)+\psi(v_i,v_j|D). \quad (7)$$

The $\varphi(v_i,v_j)$ term has the form given in the Generalized Potts model [7],

$$\varphi(v_i,v_j)=\begin{cases}K_{ij} & if \quad x_i\neq x_j\\ 0 & if \quad x_i=x_j.\end{cases} \quad (8)$$

The $\psi(v_i,v_j|D)$ term integrates the similarity of depth values for setting the labels,

$$\psi(v_i,v_j|D)=\begin{cases}\exp\left(\frac{-g^2(i,j)}{2\sigma^2}\right)\frac{1}{dist(i,j)} & if \quad x_i\neq x_j\\ 0 & if \quad x_i=x_j,\end{cases} \quad (9)$$

where $g^2(i,j)$ measures the difference in depth values of point i and j. $dist(i,j)$ denotes the Euclidean distance between i and j. σ is a constant.

Let total smooth energy $E_2(V)=\sum_{i\in S}\sum_{j\in N_i} f_{ij}(v_i,v_j|D)$, the value of V is the solution of maximizing $P(V|D)$ or equivalent minimizing the negative log-likelihood $-\log P(V|D)$

$$\min_V E_1(\Theta,V)+E_2(V)=-\log P(V|D)+const. \quad (10)$$

Model fitting is integrated by cooperating model parameter Θ into this minimizing problem. The overall problem is given by

$$\min_{\Theta,V}\phi(\Theta,V)=E_1(\Theta,V)+E_2(V). \quad (11)$$

This objective is solved using the hard EM to find the optimal values of V,Θ in an iterative progress. EM iterates the following two steps:

- **E-Step.** The model parameters are obtained from the previous iterate or from the initial step as mentioned in the previous section. Our goal is to perform the label assignment by solving $\min_V \phi(\Theta_{old},V)$ using the alpha-beta or alpha-expansion algorithm [8].
- **M-Step.** With the label assignment supplied by E-Step, we minimize the error between the model and the cloud of stereo data, $\min_\Theta \phi(\Theta,V_{new})$ by the damped least square estimator to find new parameter Θ_{new}.

III. EXPERIMENTAL RESULTS

Fig. 7 shows the reconstructed 3-D human poses from the simulated depth images. These depth images present the sequence of human walking from a front view and similar to the stereo images. The depth information is encoded in the gray color. Using binary silhouettes in the conventional way, it is impossible to compute the 3-D human configuration. With the gray silhouettes as in our case, our algorithm obtains good estimation of body configuration in 3-D that looks appropriate in the visual inspection and exactly shows the position of hands and legs in 3-D.

We have also conducted real experiments with the stereo camera (Bumblebee 2, Point Grey Research) as in Fig. 8. The Growing Correspondence Seeds (GCS) [9] algorithm is used to estimate the depth images from a pair of stereo images. Naturally, we extract the Region of Interest (ROI) correspondent to the human shape by eliminating the static background. Despite the appearance of noise and artifacts in the obtained stereo image, we can still get the model fitting result as shown in Fig. 8 (c).

ACKNOWLEDGMENT

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2008-C1090-0801-0002).

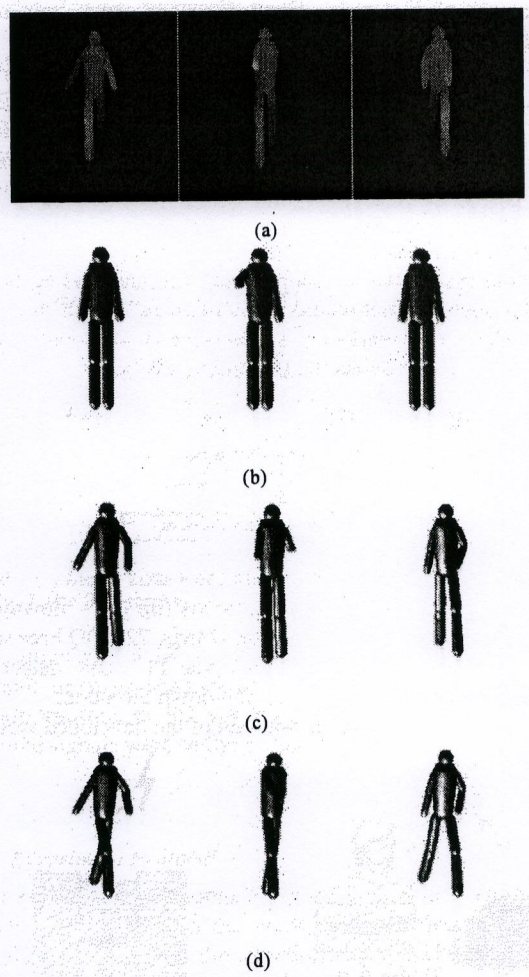


Fig. 7. Experimental results with simulation data. (a) Depth Images. (b) Initial Human Body Model. (c) Estimated and reconfigured 3-D human body model from a front view. (d) Estimated and reconfigured 3-D human body mode from a side view.

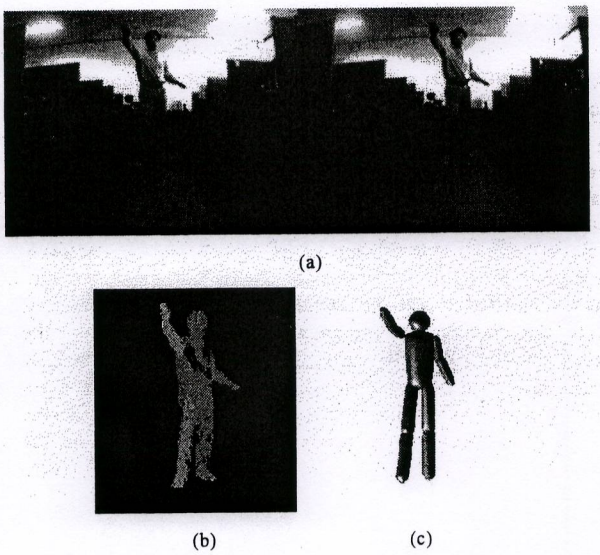


Fig. 8. Experimental results with real data. (a) A pair of stereo images obtained from a Bumblebee 2 stereo camera. (b) Depth image derived from (a). (c) Estimated 3-D human pose.

REFERENCES

[1] C. Menier, E. Boyer, B. Raffin, "3D skeleton-based body recovery," Proceedings of the Third International Symposium on 3D Data Processing, Visualization and Transmission, pp. 389-296, 2006.

[2] D. Knossow, R. Ronfard, R. Horaud, "Human motion tracking with a kinematic parameterization of extremal contours", International J. Computer Vision, vol. 79, no.3, pp. 247-269, September 2008.

[3] H. D. Yang and S. W. Lee G. O. Young, "Reconstruction of 3D human body pose from stereo image sequences based on top-down learning," Pattern Recognition vol. 40, no. 11, pp. 3120–3131, November 2007.

[4] S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Trans PAMI, vol. 24, no.24, pp. 509-522, April 2002.

[5] G. Mori, S. Belongie and J. Malik, "Efficient shape matching using shape contexts," IEEE Trans PAMI, vol. 27, no. 11, pp. 1832-1837, November 2005.

[6] T. Toyoda and O. Hasegawa, "Random field model for integration of local information and global information," IEEE Trans PAMI, vol. 30, no.8, pp. 1483-1489, August 2008.

[7] P. Kohli, J. Rihan and M. Bray, "Simultaneous segment and pose estimation of humans using dynamic graph cuts," International J. Computer Vision, vol. 79, no. 3, pp. 285-298 2008.

[8] Y. Boykov, O. Veksler and R. Zabih, "Fast approximate energy minimization via graph cuts,", IEEE Trans PAMI, vol. 23, no. 11, pp. 1222-1239, November 2001.

[9] J. Cech and R. Sara, "Efficient sampling of disparity space for fast and accurate matching" Proc. BenCOS workshop CVPR 2007.

[10] A. Sundaresan, R. Chellappa, "Model driven segmentation of articulating humans in Laplacian eigenspace", IEEE Trans PAMI, vol. 30, no. 10, pp. 1771-1785, October 2008.