Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 3120–3131
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

# Reconstruction of 3D human body pose from stereo image sequences based on top-down learning

Hee-Deok Yang, Seong-Whan Lee*

*Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Republic of Korea*

Received 22 October 2005; received in revised form 26 December 2006; accepted 10 January 2007

**Abstract**

This paper presents a novel method for reconstructing a 3D human body pose from stereo image sequences based on a top-down learning method. However, it is inefficient to build a statistical model using all training data. Therefore, the training data is hierarchically divided into several clusters to reduce the complexity of the learning problem. In the learning stage, the human body model database is hierarchically constructed by classifying the training data into several sub-clusters with silhouette images. The data of each cluster in the bottom level is represented by a linear combination of examples. In the reconstruction stage, the proposed method hierarchically searches a cluster for the best matching silhouette image using a silhouette history image (SHI). Then, the 3D human body pose is reconstructed from a depth image using a linear combination of examples method. By using depth information to reconstruct 3D human body pose, the similar poses in silhouette images are estimated as different 3D human body poses. The experimental results demonstrate that the proposed method is efficient and effective for reconstructing 3D human body poses.


*Keywords:* Reconstruction of 3D human body pose; 3D human modeling; Depth information; Spatio–temporal features

## 1. Introduction

In every day life, humans can easily understand human body poses from low-resolution images or from images obtained directly through human vision which uses two eyes for capturing images. In effect, humans have a stereo vision system [1].

Recognizing a human body pose is one of the most difficult and commonly occurring problems in a computer vision system. Much research has been developed for reconstructing a 2D or 3D human body pose from 2D information such as edge, silhouette features extracted from a monocular image, e.g., from a photograph or a video. As presented in Fig. 1, two different human body poses are similar in the 2D silhouette images. So, it is difficult to distinguish apart the pose with the left leg forward from the pose with the right leg forward. However, even if the silhouette images of two human body poses are similar, the depth images are different from each other. Therefore, the depth information is used to overcome the ambiguous 2D features.

In the proposed method, the human body pose is represented by a linear combination of prototypes of 2D depth images and their corresponding 3D joint positions. A linear combination of examples method has been applied to solve various computer vision problems [2,20]. Ullman and Basri [2] have shown theoretically that a 3D object can be represented by a linear combination of 2D prototypes of the object. They build various 2D prototypes of an object by projecting 3D feature of the object into 2D feature. Then, the 2D feature of the 3D object is represented by a linear combination of the 2D prototypes of the object. They used edge map as the 2D features of the object. However, in this paper, the depth images are used as the 2D features of a 3D human body pose.

Various human body poses are generated using 3D human body model to build statistical models based on a linear combination of examples method. However, it is inefficient to build a statistical model using all the training data. So, in the learning stage, the human body poses are hierarchically divided into several sub-clusters to reduce the complexity of the learning

* Corresponding author : Tel.: +82 2 3290 3197; fax: +82 2 926 2168.
*E-mail addresses:* hdyang@image.korea.ac.kr (H.-D Yang), swlee@image.korea.ac.kr (S.-W. Lee).

0031-3203/$30.00 © 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.01.033

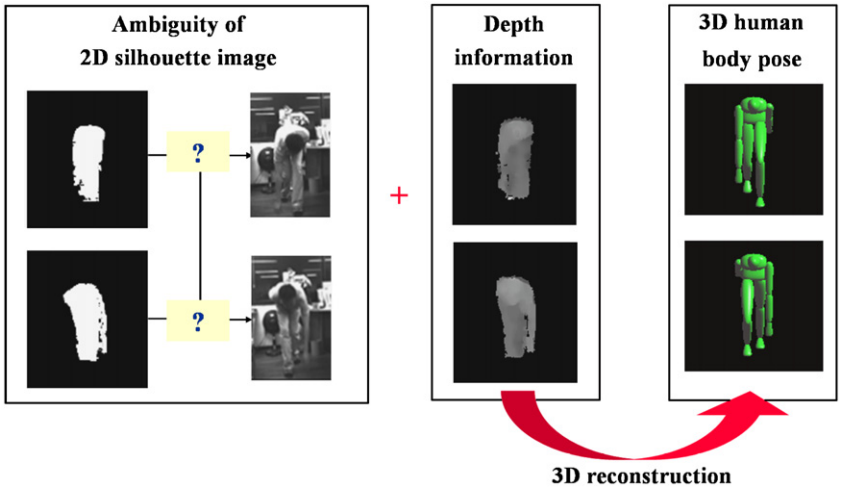


Fig. 1. Motivation of research.

problem. A blurred silhouette history image (SHI) is used as a spatio–temporal feature to reduce noise arising from the extraction of the silhouette image and also to extend the search area of a current body pose to the related body poses by accumulating the silhouette images.

The contributions of the proposed approach are: (1) the 3D human body pose database is built by hierarchically dividing training data into several sub-clusters to reduce learning complexity; (2) the spatio–temporal features are used to reduce noise occurring in the extraction of silhouette images; (3) the depth images are used to overcome the ambiguity of 2D silhouette images.

This paper is organized as follows. In Section 2, related work is briefly reviewed with a focus on reconstructing a human body pose. Section 3 describes the 3D human body model used to generate training data. Section 4 explains a 3D human body pose database and a method for 3D human body pose representation. Section 5 provides a method for the reconstruction of a 3D human body pose. Section 6 provides experimental results. Finally, Section 7 concludes this paper.

## 2. Related work

In computer vision, reconstruction of a human body pose is an interesting problem. Much research has been developed for reconstructing 2D or 3D human body poses [1,3–12,14,19,21,23]. There are two main approaches to this problem. The first is a model-based approach. This approach fits a human body model to an input image by measuring the similarity between the overlap of a human body model and an associated image region [1]. The second is a learning-based approach. This approach estimates a human body pose by searching for an image which is the best match to an input image in the training data [3]. The proposed method is based on the learning-based approach. So, several learning-based approaches are reviewed in this section.

Rosales and Sclaroff [1] used specialized mapping architecture (SMA). The SMA's fundamental comprises a set of specialized mapping functions and a single feedback matching function. Each specialized mapping function maps certain areas of the input space to the output space. The input of the algorithm is a silhouette image and the output is the reconstructed human body pose. They generated approximately 60,000 examples for training model and about 10,000 examples for testing. They used a divide-and-conquer method to reduce the complexity of the learning problem when building the specialized mapping functions using silhouette images. However, they ignored motion information that is useful for constraining the human body pose.

Bowden et al. [4] used a non-linear statistical model consisting of the positions of 2D shape contour and its corresponding 3D skeleton vertices. The 2D features are fused with the 3D features and analyzed using a point distribution model (PDM). A PDM allows a direct mapping between the positions of 2D shape contour and its corresponding 3D skeleton vertices. Their method is needed a precise silhouette extraction method to extract 2D shape contour.

Howe et al. [8] used a statistical approach by applying a Gaussian probability model for a short human motion sequence. They grouped 11 successive frames into one short motion element called a snippet. They clustered the snippets into $m$ groups using $k$-means clustering. The clustering is used to build a Gaussian probability model. They estimated human body pose by computing prior probabilities of different 3D motions for an input snippet using the trained Gaussians probability model. They assumed that 2D tracking of joint positions was provided to estimate the motion of an object.

Shakhnarovich et al. [11] developed an algorithm which used hashing-based search techniques to find relevant examples in a large database. They implicitly assumed independence between the features when estimating the parameters of hash functions. They used a locality-sensitive hashing method and locally-weighted regression for estimating the parameters. They improved the reconstruction results of 3D human body poses when compared with global hashing methods by using a locality-sensitive hashing. They generate approximately 1,775,000 examples.

Unlike these previous methods, our method considers motion and shape at the same time and also considers depth information to reduce the ambiguity of 2D features.

## 3. 3D human body model

The 3D human body modeling includes human body representation and kinematics. For human body representation, a 3D human body model which consists of body components and joints is built with perfect coordination among them. For human body kinematics, simple forward and inverse kinematics is built.

### 3.1. Human body representation

The 3D human body model consists of 17 body components. Fig. 2 shows the structure of the proposed 3D human body model. The human body model has 40 degrees of freedom (DOF). Thirty-seven DOF is used for the body model and three DOF is used for global translation. Table 1 shows the DOF of each body component.

Regarding the shape, it is felt that simple cylindrical primitives would not accurately represent body components such as the head and torso. Therefore, tapered superquadrics are employed. Superquadrics are a family of parametric shapes capable of modeling a large set of blob-like objects, such as spheres, cylinders, and parallelepiped shapes [13–15,18,22,24,25]. The superquadrics surface is given such as

$$X(\eta,\omega)=\begin{pmatrix}x(\eta,\omega)\\ y(\eta,\omega)\\ z(\eta,\omega)\end{pmatrix}=a\begin{pmatrix}a_1C_\eta^{\varepsilon_1}C_\omega^{\varepsilon_2}\\ a_2S_\eta^{\varepsilon_1}\\ a_3C_\eta^{\varepsilon_1}S_\omega^{\varepsilon_2}\end{pmatrix}, \tag{1}$$

where $S_\eta^\varepsilon=\operatorname{sgn}(\sin\eta)|\sin\eta|^\varepsilon$, $C_\eta^\varepsilon=\operatorname{sgn}(\cos\eta)|\cos\eta|^\varepsilon$, $-\pi/2\leqslant\eta,\omega\leqslant\pi/2$, the parameters $a_1$, $a_2$, and $a_3$ are the size of superquadrics along the $x$-, $y$-, and $z$-axis, respectively, $\varepsilon_1$ and $\varepsilon_2$ are the shape parameters and control the round, square, and pinched shape of the body component, and $a$ is the global scale.

Physically, $a$, $a_1$, $a_2$, $a_3\geqslant 0$ define the anthropometric parameters of the body components. Furthermore, superquadrics shapes can be deformed with tapering, bending, and twisting. In the human body model, different body components are represented with superquadrics linear tapering along $y$-axis. The deformed superquadrics parametric equation is written as

$$X'(\eta,\omega)=\begin{pmatrix}\left(\frac{k_xy}{aa_2}+1\right)\cdot x(\eta,\omega)\\ y(\eta,\omega)\\ \left(\frac{k_zy}{aa_2}+1\right)\cdot x(\eta,\omega)\end{pmatrix}, \tag{2}$$

where $-1\leqslant k_x,k_z\leqslant 1$ are the tapering parameters along $x$- and $z$-axis.

The deformable human body shape $S$ is represented by the following parameter vector:

$$S=(SQ_1,\ldots,SQ_n), \tag{3}$$

where $SQ_i=(a^i,a_1^i,a_2^i,a_3^i,\varepsilon_1^i,\varepsilon_2^i,\varepsilon_3^i,k_x^i,k_z^i)$ is the deformable superquadrics parameter set for body part $i$ and $n$ is the number of body components of the 3D human body model.

### 3.2. Human body kinematics

The transformation of each body component is described by the three movements; flexion, rotation, and abduction as shown in Table 1. The local transformation of each body component is represented by Eq. (4):

$$T=(F,R,A), \tag{4}$$

where $F$ is flexion, $R$ is rotation, and $A$ is abduction.

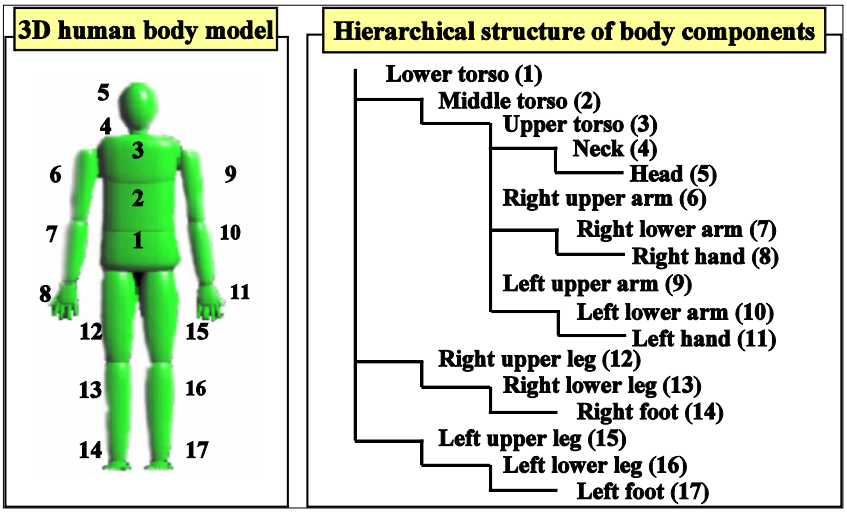


Fig. 2. The proposed 3D human model.

Table 1
DOF of each body component

| Body components | DOF | Body components | DOF | Body components | DOF |
|---|---|---|---|---|---|
| Lower torso | 3 ($F$, $R$, $A$) | Right lower arm | 1 ($F$) | Right lower leg | 1 ($F$) |
| Middle torso | 1 ($F$) | Right hand | 2 ($F$, $A$) | Right leg | 2 ($F$, $A$) |
| Upper torso | 3 ($F$, $R$, $A$) | Left upper arm | 3 ($F$, $R$, $A$) | Left upper leg | 3 ($F$, $R$, $A$) |
| Neck | 3 ($F$, $R$, $A$) | Left lower arm | 1 ($F$) | Left lower leg | 1 ($F$) |
| Head | 3 ($F$, $R$, $A$) | Left hand | 2 ($F$, $A$) | Left leg | 2 ($F$, $A$) |
| Right upper arm | 3 ($F$, $R$, $A$) | Right upper leg | 3 ($F$, $R$, $A$) | | |

$F$ (flexion): movement of $x$-axis
$R$ (rotation): movement of $y$-axis
$A$ (abduction): movement of $z$-axis.

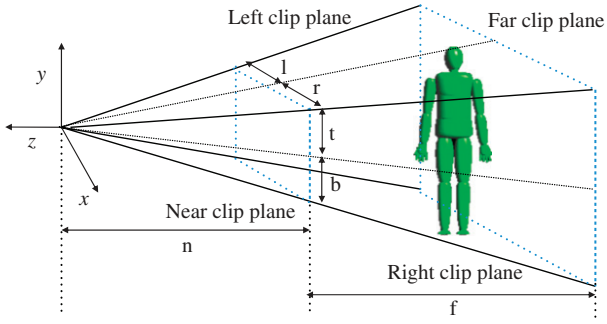


Fig. 3. Perspective camera model.

A sequence of body components is called a kinematics chain. In a kinematics chain, a child is a body component connected directly below the current body component and a grandchild is a body component connected below the child of the current body component. The transformation of a parent influences on the transformation of its child, then grandchild and so on. Therefore, if the local transformation of a parent is created, then it is concatenated with the local transformation of its descendants. The accumulated forward transformation of the body component $i$ is calculated such as

$$TA_i = T_r \ldots T_i, \tag{5}$$

where $T_r$ is the local transformation of the root and $T_j$ is the local transformation of $j$th component of the kinematics chain.

Conversely, inverse kinematics computes the joint angles of a kinematics chain based on the position of the last kinematics component. The first joint is a base and the last is an end-effector. The base joint is an anchor that cannot move, while the end-effector is the joint that can move. Inverse kinematics is used to estimate 3D human body pose using the calculated angles between joint positions. The inverse kinematics used in this paper is based on the literature [15].

### 3.3. Perspective camera model

A camera's view range is restricted by the camera's clip planes as shown in Fig. 3. The perspective camera model which describes the relation between the 3D location of object points and the 2D location of their projections is generally used to specify the camera's view range using the camera's clip planes.

The perspective camera model is also used when building the 3D human body model to approximate the behavior of a real camera. Fig. 3 presents the perspective camera model and Eq. (6) shows the perspective projection matrix. The depth information of a human body can be generated by using the perspective camera model:

$$M = \begin{pmatrix} \frac{2|n|}{(r-l)} & 0 & \frac{(r+l)}{(r-l)} & 0 \\ 0 & \frac{2|n|}{(t-b)} & \frac{(t+b)}{(t-b)} & 0 \\ 0 & 0 & \frac{|n|+|f|}{|n|-|f|} & \frac{2|n||f|}{|n|-|f|} \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{6}$$

where $n$ is the distance between a camera and a near clip plane, $f$ is the distance between a near clip plane and a far clip plane, $r$ is the distance between the projected center of a camera to the $y$-axis of a near clip plane and a right clip plane, $l$ is the distance between the projected center of a camera to the $y$-axis of a near clip plane and a left clip plane, $t$ is the distance between the projected center of a camera to the $x$-axis of a near clip plane and top of a near clip plane, and $b$ is the distance between the projected center of a camera to the $x$-axis of a near clip plane and the bottom of a near clip plane.

## 4. Hierarchical data learning

The proposed method uses a hierarchical human body pose database (HHBPDB) which is constructed by recursively classifying training data into several sub-clusters. After modeling the HHBPDB, the proposed method hierarchically searches for the best matching silhouette image from higher level to lower level of the HHBPDB. After reaching the bottom level of the HHBPDB, depth information is used to overcome the ambiguity of 2D silhouette images based on a linear combination of examples method.

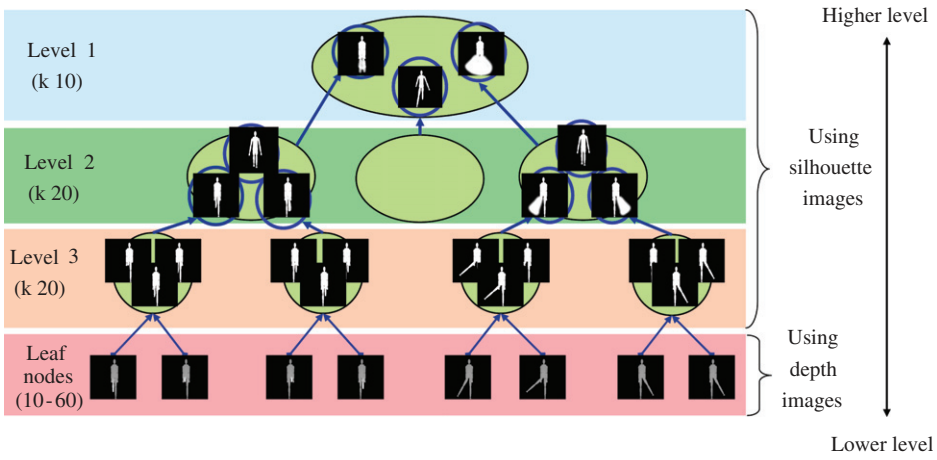


Fig. 4. Structure of the hierarchical human body pose database.

### 4.1. Data normalization

Silhouette images, depth images, and their corresponding 3D human models are generated to build the HHBPDB using the 3D human body model described in Section 3. The silhouette image and depth image are normalized to be not affected by the shapes. Let $F$ be the foreground pixels of a silhouette image:

$$F = ((x_1, y_1), \ldots, (x_{N_F}, y_{N_F}))^{\mathrm{T}}, \tag{7}$$

where $(x_i, y_i)$ is the position of the $i$th pixel of the silhouette image, $N_F$ is the number of the foreground pixels in the silhouette image.

To normalize data, a region that has $t$ percent of the foreground pixels is selected by growing pixels from the centroid of an object to outside of the object. The normalization process consists of five-steps.

*Step* 1: Apply PCA for the foreground pixels of a silhouette image to obtain eigenvectors. Rotate the silhouette image by the first eigenvector with respect to $y$-axis. Let $F'$ be the foreground pixels of the rotated silhouette image.

*Step* 2: Initialize a rectangle region $RC(C_x, C_y, L_x, L_y)$, where $C_x$ and $C_y$ are the coordinate of the center of $F'$ , $L_x$ and $L_y$ are the length of $x$- and $y$-axis, respectively. Assign 0 into $L_x$ and $L_y$.

*Step* 3: Expand the region $RC(C_x, C_y, L_x, L_y)$ using Eq. (8):

$$L_x = L_x + 2, \quad L_y = L_y + (2 \times (M_{y'}/M_{x'})), \tag{8}$$

where $M_{x'} = \max_{\{x'_i, y'_i\} \in F'}(C_x - x'_i)$ and $M_{y'} = \max_{\{x'_i, y'_i\} \in F'}(C_y - y'_i)$

*Step* 4: Calculate the number of the foreground pixels in the region $RC$ using Eq. (9):

$$N_A = \sum_{i=C_x - L_x/2}^{C_x + L_x/2} \sum_{j=C_y - L_y/2}^{C_y + L_y/2} f(x_i, y_j)$$

$$\text{where } f(x_i, y_j) = \begin{cases} 1 & \text{if } (x_i, y_j) \in F', \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

*Step* 5: If the ratio of the selected region and the foreground region satisfies Eq. (10) then stop. Otherwise, go to Step 3:

$$\frac{N_A}{N_F} \cdot 100 < t. \tag{10}$$

After selecting the region, the silhouette images and depth images are resized to have the $L_y$ with respect to fixed length $L$.

### 4.2. Hierarchical human body pose database (HHBPDB)

Several million examples are generated to build a statistical model using a linear combination method. However, it is inefficient to build a statistical model using all training data. So, the generated examples are hierarchically divided into several sub-clusters. A set of cluster is built in which each has similar shape in silhouette images. The HHBPDB has four-levels as shown in Fig. 4. To divide training data into sub-clusters, $k$-means clustering algorithm is applied.

In Fig. 4, the first, second and third levels have the mean values of clusters in the lower-level. Therefore, the number of elements of level $h$ is the number of the sub-cluster of it. In the bottom level, each cluster is represented by Eq. (11).

### 4.3. 3D human body pose representation

A linear combination method is used to reconstruct 3D human body pose from an input depth image at the bottom level of the HHBPDB. An input 2D depth image is reconstructed by a linear combination of prototypes of 2D depth image. The reconstructed 3D body model is obtained by applying the estimated coefficients to the corresponding 3D body model of the prototypes as shown in Fig. 5.

A depth image is represented by a vector $d = (d'_1, \ldots, d'_n)^{\mathrm{T}}$, where $n$ is the number of pixels in the image and $d'$ is the value of a pixel in the image. The 3D human model is represented by a vector $p = ((x_1, y_1, z_1), \ldots, (x_q, y_q, z_q))^{\mathrm{T}}$, where $x$, $y$, and $z$ are the position of body joint in the 3D space and $q$ is the number of body components of the 3D human model. Eq. (11)

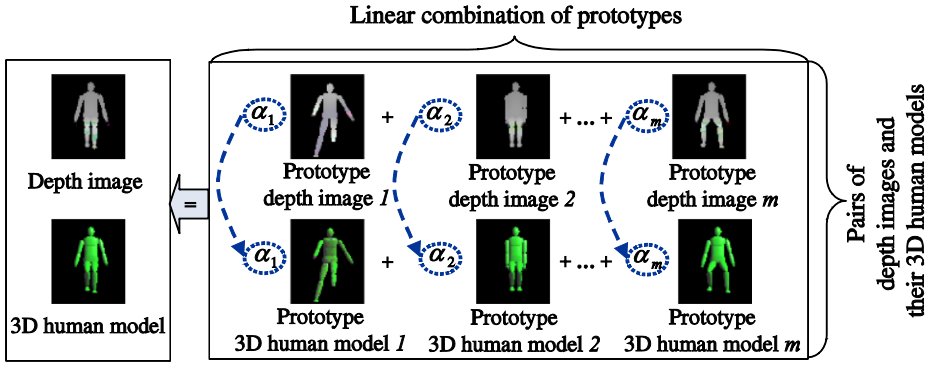


Fig. 5. Method for representing 3D human body pose.

explains training data:

$$D = (d_1, \ldots, d_m), \quad P = (p_1, \ldots, p_m), \tag{11}$$

where $m$ is the number of prototypes.

The goal is to find an optimal parameter set $\alpha$ which best estimates the 3D human body pose from a given depth image. Then, its 3D body model is represented such as

$$\tilde{D} = \sum_{i=1}^{m} \alpha_i d_i, \quad \tilde{P} = \sum_{i=1}^{m} \alpha_i p_i. \tag{12}$$

## 5. Reconstruction of 3D human body pose

A template matching method is used to find the cluster that has the best similar silhouette image in the HHBPDB with a given SHI. The reconstruction process consists of three-steps.

*Step* 1: Create a SHI and a depth image.

*Step* 2: Match a SHI image and the elements of a cluster at the level $h$, if it does not reach the leaf node of the HHBPDB. Otherwise, estimate a parameter set $\alpha$ with a depth image and reconstruct a 3D human pose with the estimated $\alpha$.

*Step* 3: Repeat Step 2 for all levels of the HHBPDB from higher level to lower level.

### 5.1. Spatio–temporal features

A temporal feature is used to reduce noise occurring in a silhouette image. A SHI which uses a method similar to a MHI [16] is applied as a temporal feature. Silhouette images are used instead of motion images to make a SHI such as

$$SHI_t(x, y) = \begin{cases} 255 & \text{if } I(x, y) = 255, \\ \max(0, SHI_{t-1}(x, y) - \lambda) & \text{otherwise,} \end{cases} \tag{13}$$

where $SHI_t(x, y)$ is the pixel at the position $(x, y)$ of the SHI at time $t$, $\lambda$ is a constant to reduce the intensity of the SHI, and $I_t$ is the silhouette image at time $t$.

The search area of the current human body pose is expanded to the related human body poses by accumulating the silhouette image.

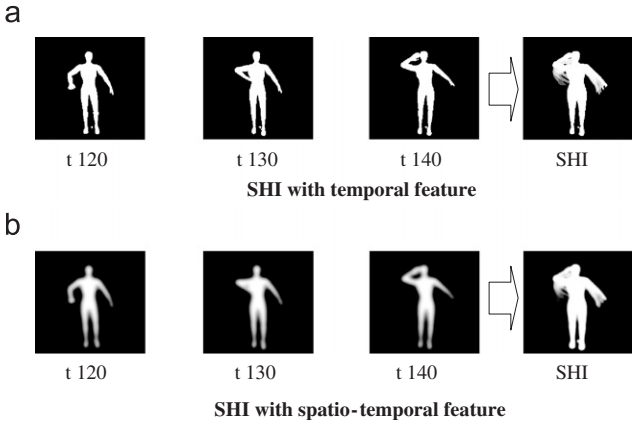


Fig. 6. Examples of temporal and spatio–temporal features.

A spatial feature is used to reduce noise. The contour of a silhouette image has noise such as a steep line, an unexpected concave or curvature. The weight of noise is reduced by blurring the silhouette image. Fig. 6 shows examples of temporal and spatio–temporal features.

### 5.2. Analyzing depth information

The depth information is extracted from stereo images. The depth image corresponding to the normalized silhouette image is selected using the method described in Section 4.1. The depth of a subject is generated such as

$$DEP = \frac{b \cdot f}{u}, \tag{14}$$

where $b$ is a baseline, $f$ is a focal length and $u$ is the distance from a subject and a focal plane of a camera.

A depth image can be normalized at a fixed distance $u$ by applying Eq. (14).

### 5.3. Matching shape

In Section 4.3, the HHBPDB is constructed to represent the training data. To search a similar silhouette image in the

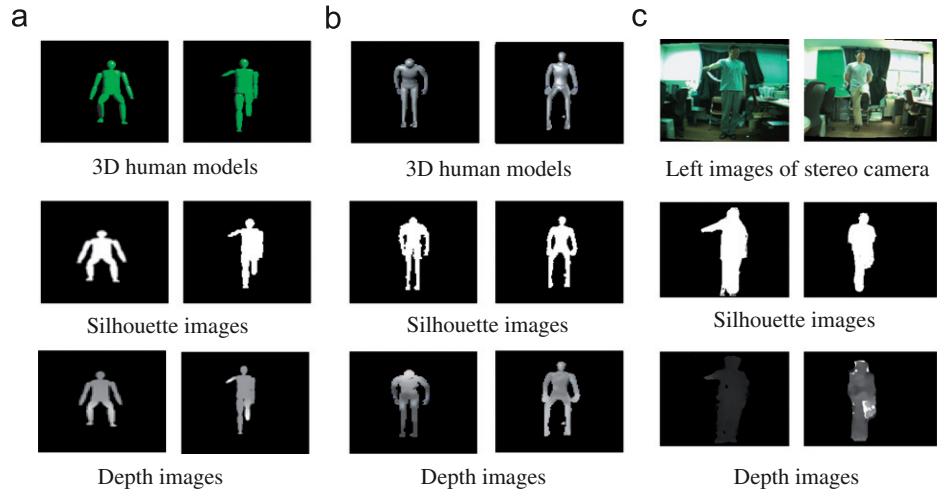


Fig. 7. Examples of training and test data: (a) training data; (b) FBG database; (c) real data.

HHBPDB, the silhouette image of each cluster of the HHBPDB and an input SHI are compared such as

$$C^h_{min} = \arg \min_{j\in\{1,\ldots,k\}} (G^h_j - SHI_t), \tag{15}$$

where $G^h_j$ is an element of the cluster at the level $h$, $SHI_t$ is the input SHI at time $t$, and $k$ is the number of elements of the cluster at the level $h$.

The search for the best matching silhouette image proceeds from higher level to lower level, until the bottom level of the HHBPDB is reached.

### 5.4. Estimating reconstruction parameter set

To reconstruct a 3D human body pose at the bottom level of the HHBPDB, the inverse matrix $D$ in Eq. (11) is calculated. The inverse $D^{-1}$ of a matrix $D$ exists only if $D$ is square. However, a matrix $D$ is not square. In this case, singular value decomposition (SVD) is applied. The pseudo-inverse of $D$ can be estimated such as

$$D = UWV^{\mathrm{T}}, \tag{16}$$

$$D^{+} = VW^{+}U^{\mathrm{T}}. \tag{17}$$

In addition, the solution $\alpha D = \tilde{D}$ can be rewritten such as

$$\alpha = D^{+}\tilde{D}. \tag{18}$$

After calculating $\alpha D = \tilde{D}$, a set of coefficients of prototypes is solved using Eq. (19). The depth of the image is calculated such as

$$\tilde{D} = \sum_{k=1}^{m} \alpha_k d_k. \tag{19}$$

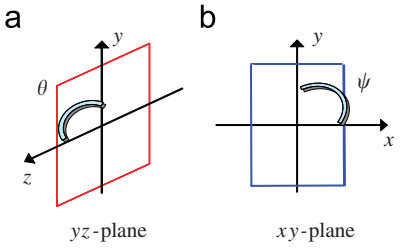


Fig. 8. Measured angles for validating: (a) $yz$-plane; (b) $xy$-plane.

Then the position of each component of the 3D human model is calculated such as

$$\tilde{P} = \sum_{k=1}^{m} \alpha_k p_k. \tag{20}$$

## 6. Experimental results and analysis

### 6.1. Experimental environment

For training the proposed method, approximately 100,000 pairs of silhouette images, depth images and their 3D human models were generated. The silhouette images are 170×190 pixels and their 3D human models are 17 joints in the 3D space. Fig. 7(a) presents the 3D human model poses in terms of the joint positions, silhouette images projected from 3D human poses at the frontal view, and depth images generated using the perspective camera model.

Two data sets were used for testing the performance of the proposed method. The first is the KU gesture database [17]. This database consists of full body gestures (FBG) database and command gesture database. The data are captured as 3D motion data and three pairs of stereo video data at three different directions for each gesture. Fig. 7(b) presents examples of the FBG database. The second consists of real data such as walking, sitting on a chair, hand raising, and bending captured with stereo

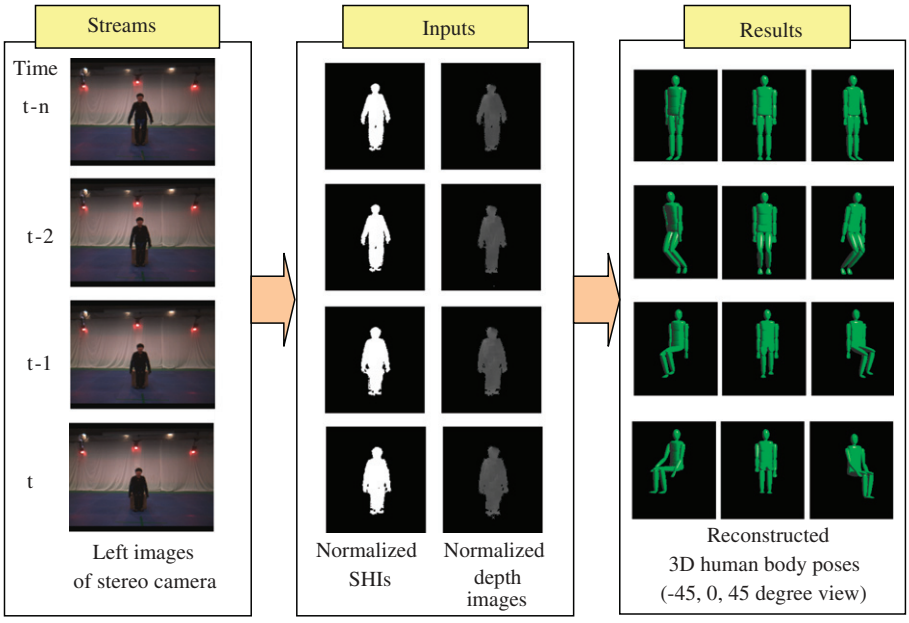


Fig. 9. Examples of the reconstructed 3D human body pose with a sitting on a chair sequence of the FBG database.

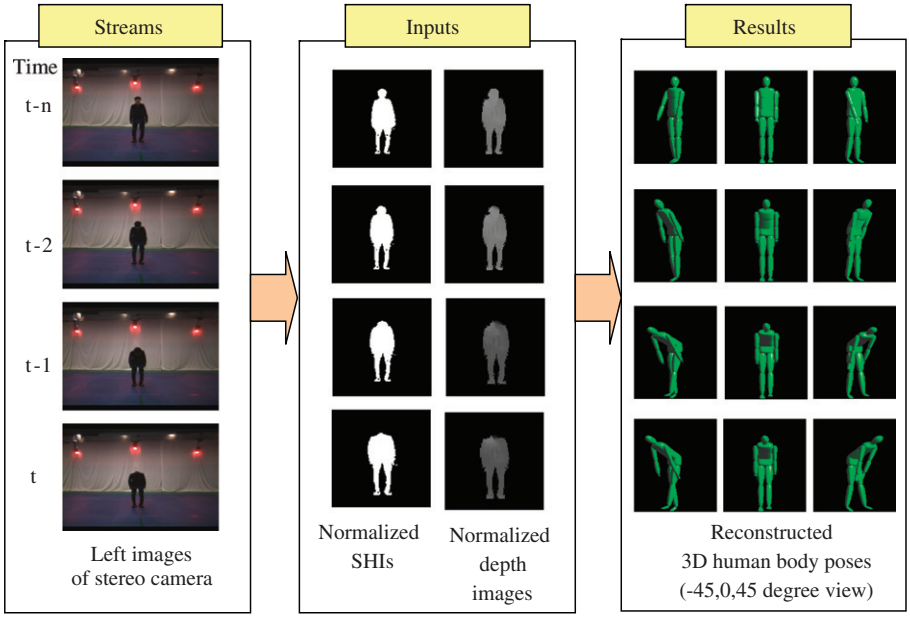


Fig. 10. Examples of the reconstructed 3D human body pose with a bending at waist sequence of the FBG database.

camera, Videre STH-MDCS2 with a resolution of 320×240. Real data was captured for two humans of their frontal and right side views simultaneously. Fig. 7(c) presents examples of real data. The silhouette images are extracted using the background subtraction method for real data and originally exists in the FBG database.

### 6.2. Experimental results

The reconstructed 3D human body poses are shown by the proposed 3D human body model using inverse kinematics. The FBG database has the ground truth for the 3D human body pose. Therefore, the ground truth and reconstructed 3D human body pose in 3D space are compared to verify the effectiveness of the proposed method. Fig. 8 shows two measured angles for verifying the proposed method. Two angles are measured in the *yz*-plane and *xy*-plane. The average error of $\theta_i$ is calculated such as

$$Err(\theta_i) = \frac{\sum_{t=1}^{n}(G(\theta_i^t) - E(\theta_i^t))}{n}, \tag{21}$$

where $n$ is the total number of frames in a sequence, $G(\theta_i^t)$ is the $\theta$ of $i$th body component at time $t$ in the ground truth and

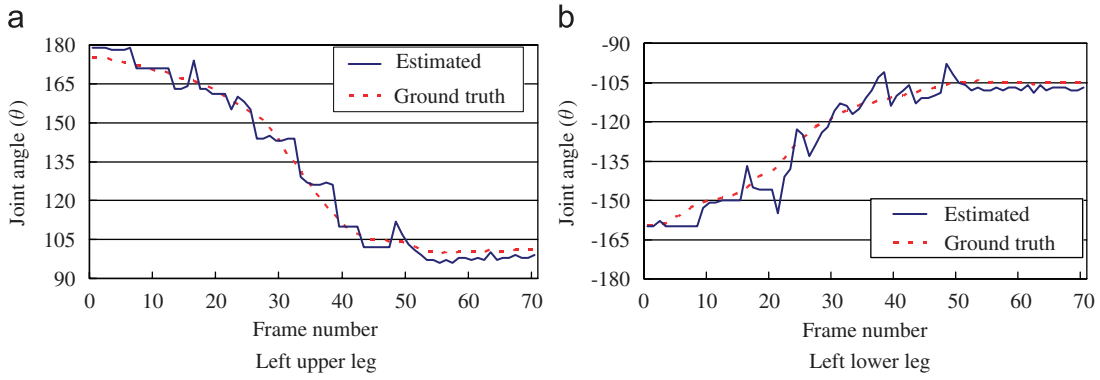


Fig. 11. Temporal curve of joint angles with the sequence in Fig. 9.

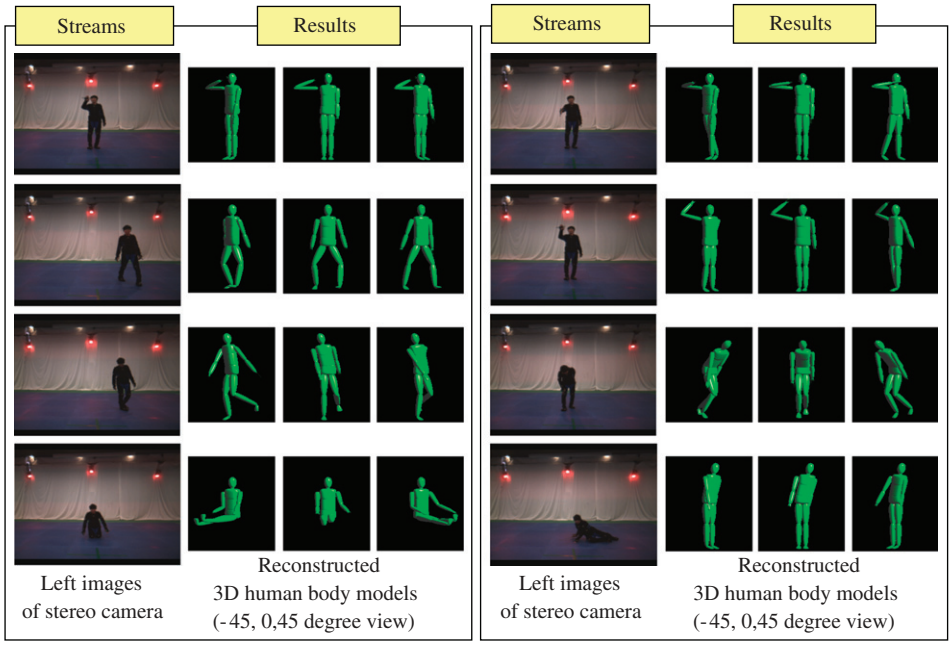


Fig. 12. Examples of the reconstructed 3D human body pose with various poses in FBG database.

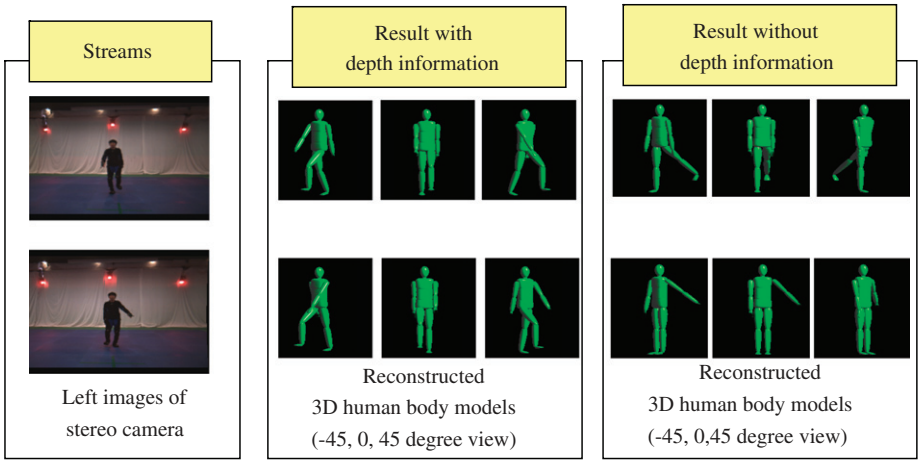


Fig. 13. Examples of the reconstructed 3D human body pose with depth information and without depth information.

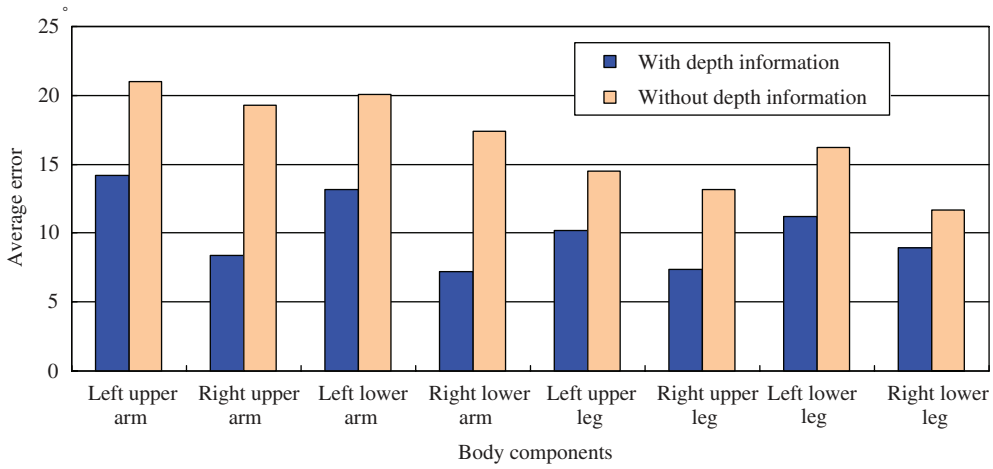


Fig. 14. The average error of joint angle with depth information and without depth information.


Fig. 15. Examples of the reconstructed 3D human body pose with various real data.

$E(\theta_i^t)$ is the $\theta$ of $i$th body component at time $t$ in the reconstructed 3D human body pose using the proposed method.

$Err(\psi_i)$ is also calculated using Eq. (21).

*6.2.1. Experiments with FBG database*

Fig. 9 shows the reconstructed results obtained using several images from the FBG database taken from the frontal view. Even though the characteristics of the human body shape differ from the characteristics used for testing, good results are achieved. The input image is a SHI and a depth image at the frontal view. The results of the estimated 3D human body model are represented a frontal view, left 45° view and right 45° view of the 3D human body model, respectively. Fig. 10 shows the reconstructed 3D human body poses with a bending at wrist sequence of the FBG database.

Fig. 11 shows the reconstructed angles of the left upper leg and the left lower leg with sitting on a chair sequence. The $\bar{\theta}$ of the left upper leg and the left lower leg are approximately 7° and 8°. The reconstructed joint angles at frames 7, 14, 27 change rapidly, because these regions are the boundaries of the clusters.

Fig. 12 shows the reconstructed 3D human body pose with various human body poses in the FBG database. The result

of the right image of the last row was rotated through the $z$-axis.

Fig. 13 shows the reconstructed results with walking in place sequence with using depth information and without using depth information. The results without using depth information contain error. The orders of hands and legs are incorrect. However, the results with using depth information show that the ambiguity occurring in the silhouette image can be resolved by using depth information. Fig. 14 shows the average error $Err(\theta_i)$ with the sequence in Fig. 13.

### 6.3. Experiments with real data

Fig. 15 shows the reconstructed 3D human body pose of various human body poses with real data. Despite low-resolution, the proposed method outputs the 3D human body pose accurately. Some results are not accurate. For example, the left image of the fifth row shows one example of a clapping of hands. The hands are located at the front of body and the depth information between hands and body is similar. Therefore, the positions of the hands are estimated incorrectly.

## 7. Conclusions and further research

In this paper, an efficient method for reconstructing a 3D human body pose from stereo image sequences is proposed using a top-down learning method. It is not efficient to build a statistical model using all examples, so the examples are hierarchically divided into several clusters. In the learning stage, the human body model database is constructed by classifying the training data recursively into several clusters with silhouette images. In the reconstruction stage, both SHI and depth images are used to reconstruct 3D human body pose. The proposed method hierarchically searches a cluster for the best matching silhouette image with an input SHI. After searching the cluster, the 3D human body pose was reconstructed with an input depth image using a linear combination method. By using depth information to reconstruct 3D human body pose, the similar poses in silhouette images can be estimated as different 3D human body poses and by using top-down learning method, the proposed method can reconstruct 3D human body poses which are not found in the training data.

One problem remaining for further research is to overcome the view point problem, current method is only suitable for the frontal view of an object. Using additive low-level information such as color, edge information and tracking extracted body component, the relationship between human body components can be analyzed.

## Acknowledgments

The authors are grateful to anonymous reviewers for helpful comments that have improved the article. This research was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-041-D00724).

## References

[1] R. Rosales, S. Sclaroff, Specialized mapping and the estimation of human body pose from a single image, Proceedings of IEEE Workshop on Human Motion, TX, USA, December 2000, pp. 19–24.
[2] S. Ullman, R. Basri, Recognition by linear combinations of models, IEEE Trans. Pattern Anal. Mach. Intell. 13 (10) (1991) 992–1006.
[3] A. Agarwal, B. Triggs, 3D human pose from silhouette by relevance vector regression, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington DC, USA, July 2004, pp. 882–888.
[4] R. Bowden, T.A. Mitchell, M. Sarhadi, Non-linear statistical models for 3D reconstruction of human pose and motion from monocular image sequences, Image Vision Comput. 18 (9) (2000) 729–737.
[5] M. Brand, Shadow puppetry, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kerkyra, Greece, 1999, pp. 1237–2344.
[6] D.M. Gavrilla, L.S. Davis, Towards 3-D model-based tracking and recognition of human movement: a multi-view approach, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 1996, pp. 73–80.
[7] T. Heap, D. Hogg, Improving specificity in PDMs using a hierarchical approach, Proceedings of the Eighth British Machine Vision Conference, Colchester, UK, September 1997, pp. 590–599.
[8] N. Howe, M. Leventon, M. Freeman, Bayesian reconstruction 3D human motion from single-camera video, Adv. Neural Inf. Process. Syst. 12 (2000) 820–826.
[9] G. Mori, et al., Recovering human body configurations: combining segmentation and recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington DC, USA, 2004, pp. 326–333.
[10] E.J. Ong, S. Gong, A dynamic human model using hybrid 2D–3D representations in hierarchical PCA space, Proceedings of 10th British Machine Vision Conference, Nottingham, UK, September 1999, pp. 33–42.
[11] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter sensitive hashing, Proceedings of Ninth International Conference on Computer Vision, Nice, France, October 2003, pp. 750–757.
[12] C. Sminchisescu, B. Triggs, Estimating articulated human motion with covariance scaled sampling, Int. J. Robotics Res. 22 (6) (2003) 371–393.
[13] A.H. Barr, Global and local deformations of solid primitives, SIGGRAPH, 1984, pp. 21–30.
[14] Y. Song, X. Feng, P. Perona, Towards detection of human motion, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, SC, USA, June 2000, pp. 810–817.
[15] D. Tolani, A. Goswami, N. Badler, Real-time inverse kinematics techniques for anthropomorphic limbs, Graphical Models 62 (5) (2000) 353–388.
[16] A. Bobick, J. Davis, The representation and recognition of action using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.
[17] B.-W. Hwang, S. Kim, S.-W. Lee, Full-body gesture database for analyzing daily human gestures, Proceedings of the First International Conference on Intelligent Computing, Hefei, China, August 2005, pp. 611–620, The KU Gesture Database, ⟨http://GestureDB.korea.ac.kr/⟩.
[18] Y. Chen, J. Lee, R. Parent, R. Machiraju, Markerless monocular motion capture using image features and physical constraints, Computer Graphics International 2005, Stony Brook, USA, June 2005, pp. 36–43.
[19] I. Infantino, A. Chella, H. Dindo, I. Macaluso, A cognitive architecture for robotic hand posture learning, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 35 (1) (2005) 42–52.
[20] M. Jones, T. Poggio, Model-based matching of line drawings by linear combination of prototypes, Proceedings of the IEEE Fifth International Conference on Computer Vision, Cambridge, UK, 1995, pp. 531–536.
[21] S.U. Lee, I. Cohen, 3D hand reconstruction from a monocular view, Proceedings of the 17th International Conference on Pattern Recognition, 2004, pp. 310–313.

[22] F. Solina, R. Bajcsy, Recovery of parametric models from range images: the case for superquadrics with global deformation, IEEE Trans. Pattern Anal. Mach. Intell. 12 (2) (1990) 131–147.

[23] C. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Comput. Vision Image Understanding 80 (3) (2000) 349–363.

[24] D. Terzopoulos, D. Metaxas, Dynamic 3D models with local and global deformations: deformable superquadrics, IEEE Trans. Pattern Anal. Mach. Intell. 13 (7) (1991) 703–714.

[25] Y. Wu, T.S. Huang, Hand modeling analysis and recognition, IEEE Signal Processing Magazine 18 (3) (2001) 51–60.

**About the Author**—HEE-DEOK YANG received his B.S. degree in Computer Science from Chungnam National University, Daejeon, Korea, in 1998; and the M.S. degree in Computer Science and Engineering from Korea University, Seoul, Korea, in 2003. Currently, he is working toward the Ph.D. degree in the Department of Computer Science and Engineering, Korea University. His research interests include sign language recognition, gesture recognition, and face recognition.

**About the Author**—SEONG-WHAN LEE received his B.S. degree in Computer Science and Statistics from Seoul National University, Seoul, Korea, in 1984, and his M.S. and Ph.D. degrees in Computer Science from KAIST in 1986 and 1989, respectively. From February 1989 to February 1995, he was an assistant professor in the Department of Computer Science at Chungbuk National University, Cheongju, Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Seoul, Korea, as an associate professor, and he is now a full professor. He was the winner of the Annual Best Paper Award of the Korea Information Science Society in 1986. He obtained the First Outstanding Young Researcher Award at the Second International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Professor Award from Chungbuk National University in 1994. He also obtained the Outstanding Research Award from the Korea Information Science Society in 1996. He also received an Honorable Mention of the Annual Pattern Recognition Society Award for an outstanding contribution to the Pattern Recognition Journal in 1998. He is a fellow of International Association for Pattern Recognition, a senior member of the IEEE Computer Society and a life member of the Korea Information Science Society. He has published more than 200 publications in these areas in international journals and conference proceedings, and has authored 10 books. His research interests include pattern recognition, computer vision and neural networks.