Human Activity Recognition Using Body Joint-Angle Features and Hidden Markov Model

Md. Zia Uddin, Nguyen Duc Thang, Jeong Tai Kim, and Tae-Seong Kim

This paper presents a novel approach for human activity recognition (HAR) using the joint angles from a 3D model of a human body. Unlike conventional approaches in which the joint angles are computed from inverse kinematic analysis of the optical marker positions captured with multiple cameras, our approach utilizes the body joint angles estimated directly from time-series activity images acquired with a single stereo camera by co-registering a 3D body model to the stereo information. The estimated joint-angle features are then mapped into codewords to generate discrete symbols for a hidden Markov model (HMM) of each activity. With these symbols, each activity is trained through the HMM, and later, all the trained HMMs are used for activity recognition. The performance of our joint-angle–based HAR has been compared to that of a conventional binary and depth silhouette-based HAR, producing significantly better results in the recognition rate, especially for the activities that are not discernible with the conventional approaches.

Keywords: Body-joint-angle features, hidden Markov model (HMM), human activity recognition (HAR).

Manuscript received June 2, 2010; revised Jan. 17, 2011; accepted Feb. 7, 2011.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (No. 2010-0001860).

Md. Zia Uddin (email: ziauddin@inha.ac.kr) was with the Department of Biomedical Engineering, Kyung Hee University, Yongin, Gyeonggi-do, Rep. of Korea, and is now with the Department of Electronic Engineering, Inha University, Incheon, Rep. of Korea.

Nguyen Duc Thang (email: ducthang@oslab.khu.ac.kr) is with the Department of Computer Engineering, Kyung Hee University, Yongin, Gyeonggi-do, Rep. of Korea.

Jeong Tai Kim (email: jtkim@khu.ac.kr) is with the Department of Architectural Engineering, Kyung Hee University, Yongin, Gyeonggi-do, Rep. of Korea.

Tae-Seong Kim (corresponding author, email: tskim@khu.ac.kr) is with the Department of Biomedical Engineering, Kyung Hee University, Yongin, Gyeonggi-do, Rep. of Korea.

doi:10.4218/etrij.11.0110.0314

I. Introduction

Human activity recognition (HAR) is defined as the recognition of different human activities utilizing external sensors, such as motion, acceleration, or video sensors. In recent years, HAR from video has evoked considerable interest among researchers of computer vision and image processing [1]-[5]. A key reason for this is the use of the outcomes of such recognition in practical applications, such as smart home, human computer interaction, automated surveillance, and human healthcare applications. For instance, an HAR system can be used at home to recognize a subject's daily activities automatically based on which a medical doctor can analyze the history of various activities over a period of time to evaluate the condition of a subject's health, which can be helpful for a better diagnosis and treatment. A general method for HAR starts with the extraction of key features and comparing them against the features of various activities. Thus, activity feature extraction, modeling, and recognition techniques become essential elements in this regard. Generally, HAR is a challenging task as it does not follow rigid syntax-like gesture or sign language recognition. Thus, a complete representation of a human body is necessary to characterize human movements properly.

So far, 2D binary silhouettes are the most popular representations of human body that have been applied for HAR [1]-[5]. For instance, in 1991, Yamato and others utilized binary silhouettes followed by vector quantization and a hidden Markov model (HMM) to recognize some time-sequential tennis activities [1]. In 2002, Carlsson and Sullivan proposed a shape matching key-frame–based approach to recognize forehand and backhand strokes from tennis video clips in which they utilized the Canny edge detector to represent the shapes [2]. In 2004, the authors utilized principal component

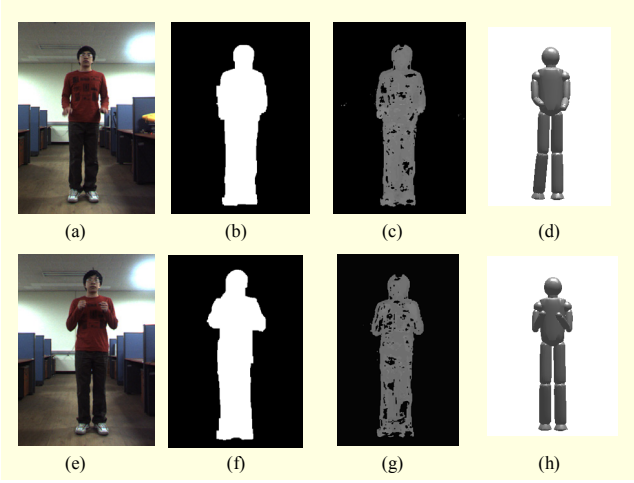

Fig. 1. (a) Sample RGB image, (b) corresponding binary silhouette, (c) depth silhouette, and (d) co-registered 3D body model from a both hands up-down activity as well as (e) sample RGB image, (f) corresponding binary silhouette, (g) depth silhouette, and (h) co-registered 3D body model from a boxing activity.

(PC) features from binary silhouettes and optical-flow–based motion features in combination with an HMM to recognize different view-invariant activities [3]. In 2009, Uddin and others proposed independent component (IC) features of binary silhouettes to recognize five different activities by means of an HMM that showed the superiority of the IC-based local features over the PC-based global silhouette features [4]. Although binary silhouettes are very commonly employed to represent a wide variety of body configurations, they produce ambiguities by representing the same silhouette for different postures from different activities. That is to say, if a person performs hand movement activities in the direction of the camera, a different posture can correspond to the same silhouette as its two-level (that is, white or black) flat pixel intensity distribution. One clear example is shown in Fig. 1 in which Figs. 1(a) and (b) represent an RGB frame and its corresponding binary silhouette, respectively, from boxing and Figs. 1(e) and (f) an RGB and its corresponding binary silhouette from both hands up-down activity. It is obvious that the binary silhouettes do not seem to be a good choice to separate these two different postures. Also, from the binary silhouettes, it is impossible to obtain the difference between the far and near parts of human body in the activity video. To improve the silhouette representation, Uddin and others proposed IC features from the time-sequential activity depth silhouettes to be used with HMMs for robust human activity recognition [5]. Although depth silhouettes are better than binary silhouettes as shown in Figs. 1(c) and (g), they still present some ambiguities that cannot be resolved. However, as the human body consists of limbs and the joints inside them, if one can get the 3D joint-angle information, then one can form much stronger features than depth and binary features, which will lead to significantly improved HAR even for activities that are not recognizable with the conventional methods.

As mentioned, to overcome the limitations of the binary and depth silhouettes, human activity recognition works more efficiently by deriving and using joint-angle information from a 3D whole body model. Conventionally, 3D human body configuration is captured by multiple cameras in a predefined environment where optical markers are placed on a subject to capture the positions of body limbs in 3D [6]. The body joint angles are then extrapolated from the 3D coordinates of the located markers through inverse kinematic analysis. However, the marker-based system is restrictive to users and cannot be used for daily activity monitoring. That is why we try to develop a marker-free HAR system in this study. Lately, capturing 3D human body configuration from a sequence of images without markers is getting considerable attention. Monocular image-based approaches have been devised in this respect as in [7]-[9] where the authors formulated complex probabilistic relationships between the parameters of 3D human models and the likelihood using image features, such as edges, contours, and silhouettes. The model parameters were then found to be the ones most probable in the given likelihood. Although these monocular image-based approaches achieve considerable speed and accuracy in obtaining 3D human body postures, there is one inherent limitation of these approaches, that is, depth ambiguity. Basically with monocular images, some depth information becomes missing to recover the original object. This makes the monocular image-based approaches ill-posed and presents a great challenge.

To overcome the disadvantages of monocular images in the 3D pose estimation, multiview approaches have been proposed where each image was processed for modeling, and finally, the outcomes were combined to obtain accurate 3D model parameters [10], [11]. However, this kind of approach requires a specialized setup in which multiple cameras are installed at different view angles to capture comprehensive 3D information. In general, deploying multiple cameras in a given location is not flexible. Also, it requires a complicated calibration process to synchronize the cameras, making this approach less practical for a daily use of HAR. These approaches have been mainly investigated for 3D pose estimation, not for HAR.

To overcome the limitations mentioned above, Thang and others developed a method from which one can estimate the body joint angles from time-series pairs of stereo images (from only a single stereo camera) without attaching any markers to a human subject and without deploying multiple cameras [12], making the approach practical to be used in daily life. In this

approach, by co-registering a 3D human body model to the stereo information (that is, depth information), 3D body postures are recovered from each pair of stereo images in time, yielding the joint angles of each body joint. Two sets of exemplars of our body posture representations are shown in Figs. 1(d) and (h) where each 3D human body configuration makes a clear separation of the two different postures, and each body segment, represented as one ellipsoid, is connected as a joint.

In this work, we report a novel method of HAR based on the body joint angles directly estimated from a pair of activity stereo images by co-registering a 3D body model to the stereo data. We have made several improvements on the original algorithm [12] to gain more speed and accuracy. To make the joint-angle features more robust, they are classified by linear discriminant analysis (LDA). Then, through a vector quantization process, the time-sequential joint-angle features are then mapped into discrete symbols to train each HMM per activity. Finally, each trained activity HMM is used to recognize each human activity. Our work focuses on HAR using the joint angles estimated without the markers and multiple cameras, yet with just a single camera. The feasibility of performing HAR based on this simple setup can be critical in practical applications, such as daily human activity monitoring without restrictions on a user. The average processing time of our proposed joint-angle–based HAR system takes 0.89 seconds per activity frame (at the resolution of 640×480) for co-registration and recognition (but 3.71 seconds per frame in depth computation on a Pentium IV of 3 GHz and 1 GB RAM), showing its feasibility for real-time HAR in the near future.

The organization of the rest of this paper is as follows. We introduce the methodology in section II where the architecture of proposed system is elaborated from stereo image preprocessing to activity recognition by an HMM. We proceed to sections III and IV to explain experimental setups and results with some discussions, respectively. Finally, we give a conclusion in section V.

II. Methodology

Our HAR system consists of the following steps: (i) stereo image processing, (ii) 3D human body modeling, (iii) joint-angle estimation, (iv) feature representation, and (v) training of HMMs for recognition. Figure 2 shows the overall architecture of our proposed activity recognition system.

1. Stereo Image Processing

To capture 3D information, we have utilized a stereo camera

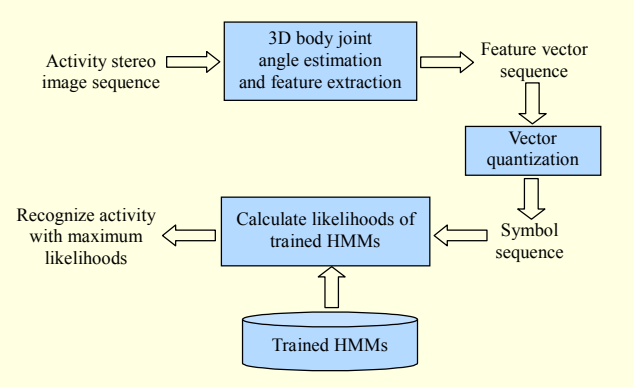

Fig. 2. Proposed human activity recognition system.

(Bumblebee 2.0 of Point Grey Research) and captured a pair of stereo RGB images at the same time in a manner similar to human eyes. The growing correspondence seeds (GCS) algorithm [13] is applied to obtain the displacements of an image pair to produce a disparity image. Then, the depth value Z of each point in 3D is computed from the disparity image by

$$Z = \frac{fb}{D}, \tag{1}$$

where f is the focus length, b is the baseline, and D is the disparity value. The two remaining coordinates X and Y are given by

$$X = \frac{uZ}{f}, \tag{2}$$

$$Y = \frac{vZ}{f}, \tag{3}$$

where u and v are the column and row index of the pixel in the disparity image, respectively.

2. 3D Human Body Modeling

Our articulated human body model is depicted in Fig. 3, where each segment of the body model is controlled by a series of the transformations specified by kinematic parameters.

At each joint, there are two degrees of freedom (DOF) that determine the transformation from the current segment to the next segment. The transformation from the global coordinate system to the local coordinate system attached at the body's hip requires six DOF (that is, three translations and three rotations). For simplification, we have utilized a limited number of the kinematic parameters, but this is enough to distinguish the main characteristics of each human posture.

There are a total of 14 body segments, 9 joints (that is, two

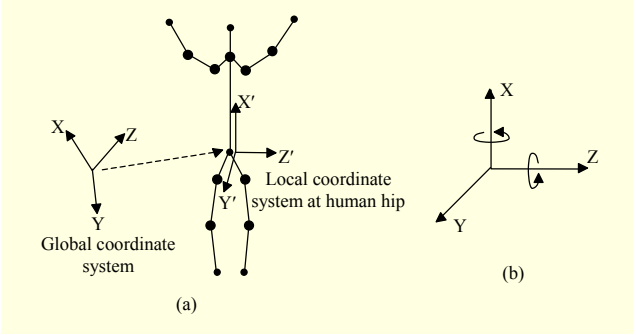

Fig. 3. (a) Articulated skeletal human body model and (b) two rotational Euler angles around the *x*-axis and *z*-axis (two DOF at each joint) controlling the movement of each segment.

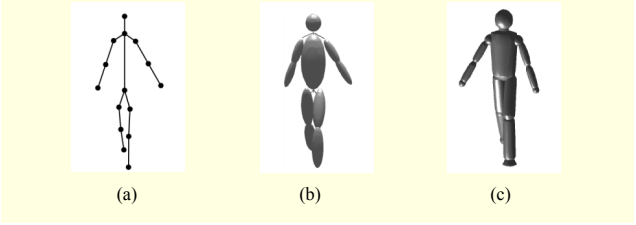

Fig. 4. Sample of (a) skeletal model, (b) computational model with ellipsoids, and (c) synthetic model with super-quadrics.

knees, two hips, two elbows, two shoulders, and one neck), and 24 DOF (that is, two DOF to the horizontal and vertical direction, respectively, at each joint and six DOF for the transformation from the global coordinate system to the local coordinate system at the body's hip). All of the 24 DOF represent the kinematic parameter $\theta = (\theta_1, \theta_2, ..., \theta_{24})$ of the human model. In addition, another synthetic human model using the super quadric is also introduced here for displaying the estimated human postures. The formulation of the super-quadric surface is introduced in [12]. Figure 4 shows the samples of a skeletal model, computational model with ellipsoids, and human model with super-quadrics.

3. Joint-Angle Estimation

In our approach of joint-angle estimation, we first define our 3D human model with a set of connected ellipsoids which are parameterized by kinematic angles. The angular kinematic angles are adjusted to fit the 3D model to the observation. Consequently, we can reconstruct human posture reflected in stereo images. Figure 5 shows the basic steps to estimate the joint-angle features, thus obtaining 3D human body posture. The first step to estimate the 3D data from depth images has been presented in the previous section. In the second step, we have included a tracking algorithm to locate the position of a moving subject. In previous work [12], the authors developed a

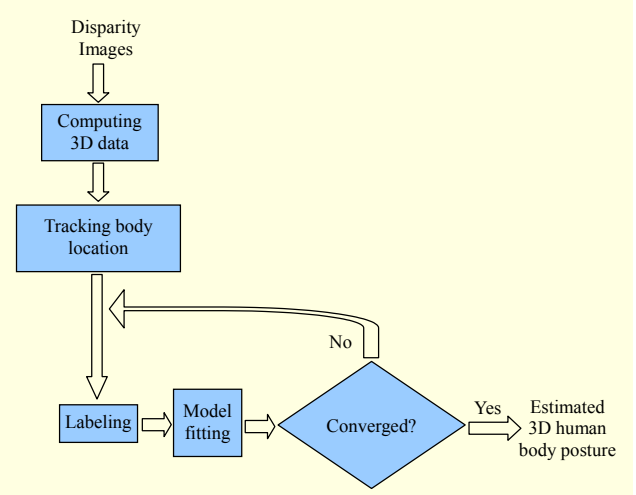

Fig. 5. Basic steps of the proposed body-joint-angle estimation.

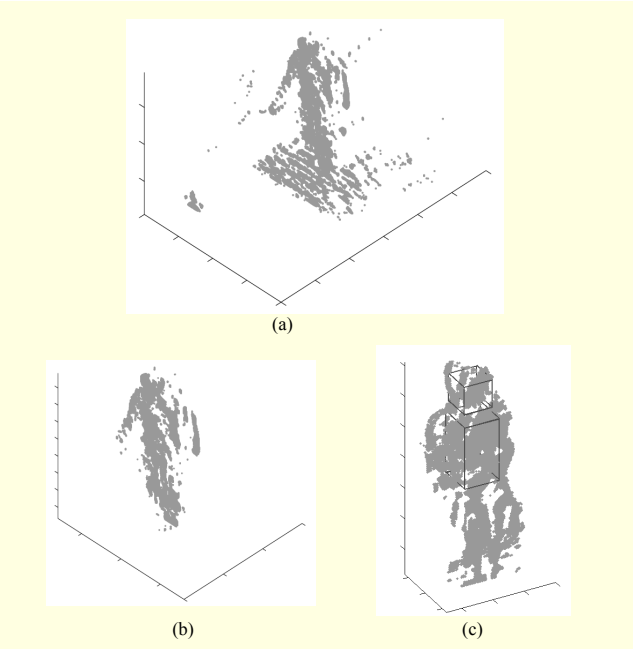

Fig. 6. Sample of (a) 3D data of moving person, (b) noise removal of 3D data of moving subject, and (c) detecting head and torso of sitting person.

system to estimate human body postures of a subject performing activities in a fixed location, making this step unnecessary. In this work, concerned with recovering human body postures of a subject moving in the horizontal (for example, walking) and vertical direction (for example, sitting), we have added the tracking step to locate the subject's position. In addition, the subject's location is used to remove the artifacts, which are a part of 3D data remaining far from the subject as depicted in Fig. 6(b). Furthermore, face detection is utilized to detect the head and torso areas as depicted in Fig. 6(c), which

are used in the labeling step of the co-registration algorithm. Finally, the six parameters of the global transformation from the global coordinate system to the local coordinate system at the body's hip are computed with the subject's location obtained by the tracking step, giving higher precision.

Let a pair of parameters $[s_t^{\mathrm{H}}, s_t^{\mathrm{B}}]$ present the location of a human subject where s_t^{H} and s_t^{B} are the two 3D vectors locating the center of the head and the body at the time index *t*. From the information of RGB images and 3D data, we can obtain the approximate values of $[s_t^{\mathrm{H}}, s_t^{\mathrm{B}}]$ by $[r_t^{\mathrm{H}}, r_t^{\mathrm{B}}]$: we detect the head region from RGB images and 3D data by the face detection algorithm using the Haar features [14] to compute r_t^{H}, and we track the body region from RGB images and 3D data using the mean shift algorithm [15] to get a value of r_t^{B}. Let $s_t'^{\mathrm{H}}$ and $s_t'^{\mathrm{B}}$ be the velocity of the head and body at the time index *t*. A set of equations established to track the changes from $[s_{t-1}^{\mathrm{H}}, s_{t-1}^{\mathrm{B}}]$ to $[s_t^{\mathrm{H}}, s_t^{\mathrm{B}}]$ and the relationship between the real human location $[s_t^{\mathrm{H}}, s_t^{\mathrm{B}}]$ and the raw estimation $[r_t^{\mathrm{H}}, r_t^{\mathrm{B}}]$ is given by

$$s_t^{\mathrm{H}} = s_{t-1}^{\mathrm{H}} + s_{t-1}'^{\mathrm{H}}\tau - \upsilon_1, \tag{4}$$

$$s_t^{\mathrm{B}} = s_{t-1}^{\mathrm{B}} + s_{t-1}'^{\mathrm{B}}\tau - \upsilon_2, \tag{5}$$

$$s_t'^{\mathrm{H}} = s_{t-1}'^{\mathrm{H}} + \upsilon_3, \tag{6}$$

$$s_t'^{\mathrm{B}} = s_{t-1}'^{\mathrm{B}} + \upsilon_4, \tag{7}$$

$$r_t^{\mathrm{H}} = s_t^{\mathrm{H}} + \zeta_1, \tag{8}$$

$$r_t^{\mathrm{B}} = s_t^{\mathrm{B}} + \zeta_2, \tag{9}$$

$$d = \left\| s_t^{\mathrm{H}} - s_t^{\mathrm{B}} \right\| + \zeta_3, \tag{10}$$

where $\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4, \zeta_1, \zeta_2$, and ζ_3 are random variables drawn from a Gaussian distribution, τ is the time interval between two frames, and *d* is the constant distance between the center of the head and the center of the body. We update the current subject's position $[s_t^{\mathrm{H}}, s_t^{\mathrm{B}}]$ from the previous estimation $[s_{t-1}^{\mathrm{H}}, s_{t-1}^{\mathrm{B}}]$ and from the observation $[r_t^{\mathrm{H}}, r_t^{\mathrm{B}}]$ by extended Kalman filter [16]. The face and torso regions are estimated from $[s_t^{\mathrm{H}}, s_t^{\mathrm{B}}]$ by the method presented in [12].

In the following step, we co-register the 3D model into the 3D data. Although more details regarding the co-registration processes can be found in [12], the co-registration including labeling and model fitting can be summarized as follows. $D = (X_1, X_2, ..., X_N)$ is considered a collection *N* 3D points, and I denotes an RGB image. The supplementary variable $V = (v_1, v_2, ..., v_N)$ is introduced to determine to which part (that is, ellipsoid) of body each point should belong. Here, a probabilistic relationship between the 3D model and the observed data is presented by the posterior probability between the label *V* and the kinematic parameter θ given the 3D data *D* and the RGB image *I*:

$$P(V, \theta \mid I, D) \propto P(V)P(I \mid V)P(D \mid V)P(D \mid V, \theta). \tag{11}$$

We now sequentially define each element of (11). The smoothness prior *P*(*V*), found from the Potts model [17], presents the pairwise probabilistic relationship of each pair of 3D points. The head and torso areas detected in RGB images and 3D data by the tracking step provide extra information about the label of 3D points. This information can be obtained by the likelihood term $P(I \mid V)$. If the geodesic distance is the shortest path distance in a graph using the Dijkstra's algorithm [18], the pairwise geodesic relationship $P(D \mid V)$ establishes some geodesic distance constraints of each pair of 3D points. Two 3D points with two corresponding labels that disregard these constraints (that is, too close or too far) are penalized to decrease the probability $P(D \mid V)$. Finally, the reconstrution error $P(D \mid V, \theta)$ is related to the total Euclidean distances from each point to the ellipsoids corresponding to the body parts.

We can see that the most suitable posture with the observed data will correspond to the kinematic parameter θ^* that maximizes the posterior probability given in (11). The EM algorithm is used for this optimization problem with the appearance of the latent variable *V*. The proposed algorithm formulated in an EM framework is an iterative procedure with two main steps, that is, E-step and M-step:

- **E-step (labeling):** Assuming that the current value of the kinematic parameter is $\theta = \theta_{\mathrm{old}}$, the E-step estimates the label assignments by computing the distribution of $P(V \mid \theta_{\mathrm{old}}, I, D)$. The true distribution of *V*, given that θ_{old}, *I*, and *D*, is intractable to computation, so we use the variational inference method called mean field [19] to approximate $P(V \mid \theta_{\mathrm{old}}, I, D)$.
- **M-step (model fitting):** With the label assignment $P(V \mid \theta_{\mathrm{old}}, I, D)$ provided by the E-step, the M-step maximizes the term $E_{P(V \mid \theta_{\mathrm{old}}, I, D)}[\log(P(V \mid \theta_{\mathrm{old}}, I, D))]$ or equivalent to minimize the reconstruction error between the model and the cloud of 3D points that is solved by the Levenberg-Marguardt least square estimator. Since the 3D data contains a thousand 3D points, a direct estimation of the kinematic parameters from a cloud of 3D points is very complicated and time consuming. We grouped a set of 3D points with the same assigned label into a Gaussian cluster. Each Gaussian cluster is parameterized by a Gaussian

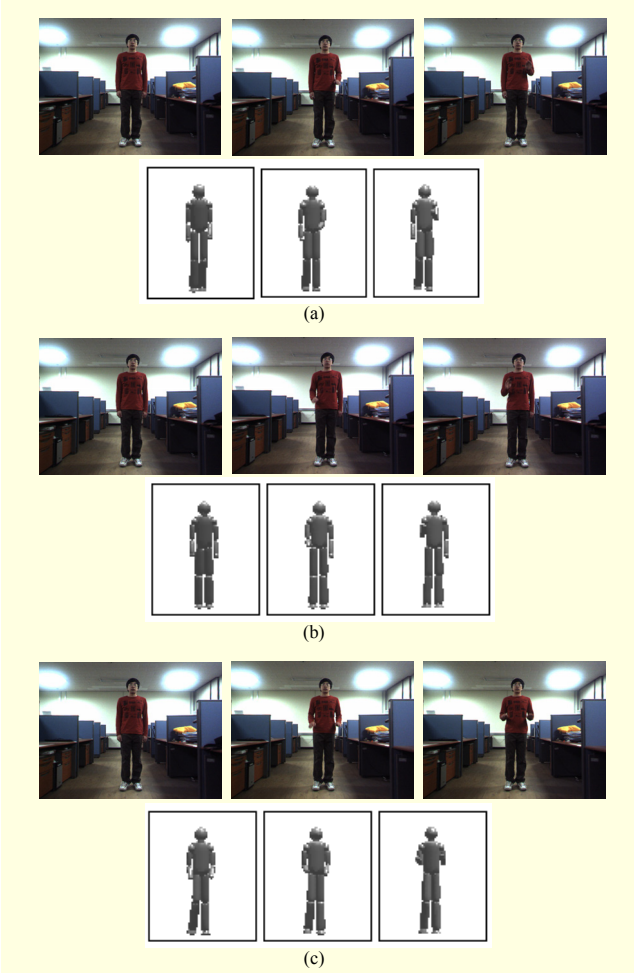

Fig. 7. Three sequential RGB images and corresponding 3D body models from (a) left hand up-down, (b) right hand up-down, and (c) both hands up-down activities.

distribution $N(m_i, \Sigma_i^{-1})$ where m_i is the center of the cluster and $\Sigma_i^{-1} = R_i^T \Lambda_{i0} R_i$ the covariance matrix. Here, the constant matrix Λ_{i0} configures the shape of the Gaussian cluster and R_i is a rotation matrix of the Gaussian cluster. The maximum likelihood is used to estimate the parameters (m_i, R_i) of each cluster. By estimating the kinematic parameters from a small number of clusters, we have consequently reduced the computational time of the co-registration algorithm, consuming less than 1 second per frame where the original approach presented in [12] took about 2 seconds for the same task.

The co-registration is iterated to minimize the differences between the 3D model and the observed data. Finally, it recovers the correct human posture with the estimated joint angles. Figure 7 shows three sequential RGB images and the corresponding 3D body models from (a) left hand up-down, (b) right hand up-down, and (c) both hands up-down activities, respectively.

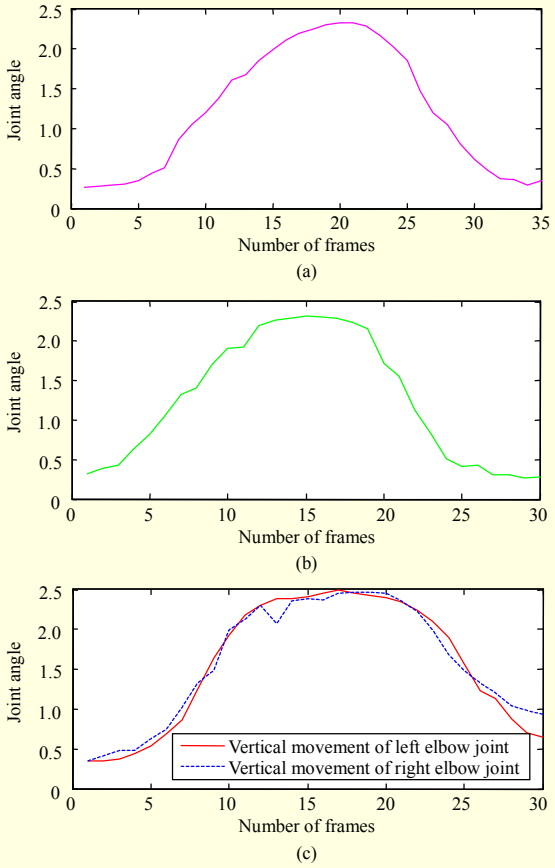

Fig. 8. Changes in vertical DOF of the joint angles of (a) left elbow, (b) right elbow, and (c) both elbows from a sequence of left hand up-down, right hand up-down, and both hands up-down activities.

4. Feature Representation

The 3D human postures are distinguished by a set of the kinematic parameters. In our work, the raw kinematic parameters, the joint angles without any adjustment, are used as the features of the human posture because these features reveal the local movements of body parts. The 24D feature vector of each human posture at time t is expressed as

$$F_t = [\theta_{\text{global\_to\_local}}, \theta_{\text{left\_shoulder}}, \theta_{\text{right\_shoulder}}, \theta_{\text{left\_crotch}}, \theta_{\text{right\_crotch}}, \theta_{\text{neck}}, \theta_{\text{left\_elbow}}, \theta_{\text{right\_elbow}}, \theta_{\text{left\_knee}}, \theta_{\text{right\_knee}}], \quad (12)$$

where $\theta_{\text{global\_to\_local}}$ consists of six DOF of the transformation from the global coordinate system to the local at the body's hip and the rest of the parameters representing a set of kinematic angles consisting of two DOF (that is, horizontal and vertical directions, respectively) at each body joint.

Figures 8(a), (b), and (c) show changes in the vertical DOF of the joint angles of left elbow, right elbow, and both elbows from a sequence of left hand up-down, right hand up-down,

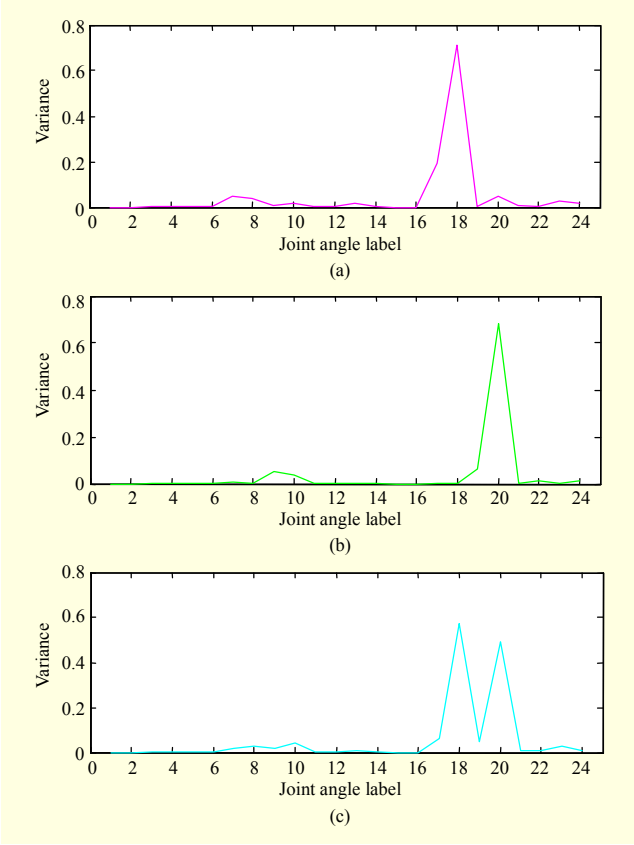

Fig. 9. Variance plot of joint angles from a sequence of (a) left hand up-down, (b) right hand up-down, and (c) both hands up-down activities.

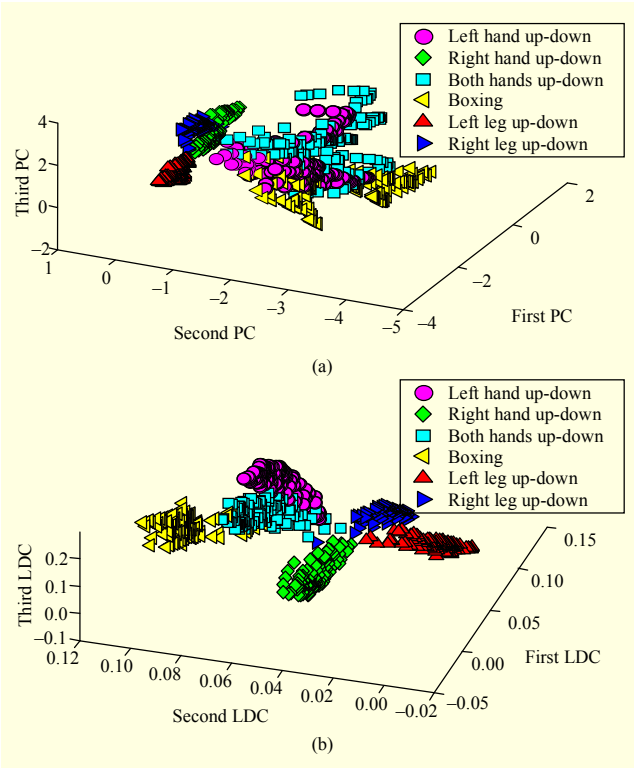

Fig. 10. 3D plot of (a) PC and (b) LDA features in local coordinate system from left hand up-down, right hand up-down, both hands up-down, boxing, left leg up-down, and right leg up-down activities.

and both hands up-down, respectively. Figure 9 represents a variance plot of joint-angle features of three different activities showing remarkable changes in the vertical DOF of the joint angles of the left elbow, right elbow, and both elbows from a sequence of left hand up-down, right hand up-down, and both hands up-down, respectively. Once the features vector for a frame is defined, we can define an activity clip as a sequence of feature vectors as $(F_1, F_2, ..., F_T)$, where T indicates the number of frames in the activity video.

For the final step of the feature extraction, LDA is performed on the angle parameters extracted from the joints of the 3D body model. Basically, LDA is based on the class specific information which maximizes the ratio of between-class scatter matrix and the within-class scatter matrix [20]. The optimal discriminant vector matrix W_{lda} is chosen from the maximization of ratio of the determinant of the between class scatter matrix S_B of the projection data to the determinant of the within class scatter matrix S_W of the projected samples as

$$J(W_{\text{lda}}) = \frac{|W_{\text{lda}}^T S_B W_{\text{lda}}|}{|W_{\text{lda}}^T S_W W_{\text{lda}}|}, \tag{13}$$

where W_{lda} is the set of discriminant vectors of S_B and S_W corresponding to the c–1 largest generalized eigenvalues.

The discriminant ratio is derived by solving the generalized eigenvalue problem such that

$$S_B W_{\text{lda}} = \Lambda S_W W_{\text{lda}} \tag{14}$$

where Λ is the eigenvalue matrix. This discriminant vectors W_{lda} forms the basis of the (c–1) dimensional subspace for a c-class problem.

Usually, the LDA algorithm looks for the vectors in the underlying space to create the best discrimination among different classes. Thus, the extracted features from 3D modeling from the images of different activities can be extended by LDA. The feature vectors using LDA on the angular joint features can be represented according to

$$L_i = F_i W_{\text{lda}}^T. \tag{15}$$

Thus, utilizing LDA on the joint-angle features, a more robust feature space can be obtained that separates the feature vectors of each class. Figure 10(b) shows the 3D plot of the LDA-based feature vectors where each activity feature is concentrated in a particular region of the feature space, indicating good separation.

5. HMMs for Activity Training and Recognition

In our work, we adopt HMM due to its capability of handling spatio-temporal features as demonstrated in some previous studies [1], [3]-[5]. An HMM can be denoted as $H = \{A, B, \pi\}$, where A denotes the state transition probability, B the observation symbol probability, and π the initial state probability. The following subsections illustrate the codebook generation to obtain discrete symbol sequences from the body-joint-angle features and activity training and recognition procedure, respectively.

A. Codebook Generation

The next step after obtaining the joint-angle features is to partition the feature space for vector quantization. We symbolize the joint-angle feature vectors from the activity frames before applying to train or recognize by the HMM. As a result, an efficient codebook of vectors must be generated using a vector quantization algorithm from the training vectors. In our work, the Linde, Buzo, and Gray (LBG) clustering algorithm [21] is used to generate a codebook from the training feature vectors. Initial selection of the centroids is obtained first. Then, until a convergence criterion is met, it finds the nearest centroid for each sample to assign it to that cluster and computes the center of all clusters after assigning all samples to the new clusters. However, the initialization is done by splitting the centroid of the whole dataset. It starts with a codeword size of one and recursively splits into two codewords. After splitting, optimization of the centroids is done to reduce the distortion. Since it follows binary splitting, the size of the codebook becomes a power of two. To obtain a symbol for a sample feature vector, the vector is compared to all the codewords and the index of one is chosen having a minimum distance. Figure 11 shows the basic steps to produce a codebook through LBG and symbol selection, that is, cluster, using the codebook.

Figure 12 shows the codeword index patterns for a sample testing feature vector sequence of different activities where all the sequences follow separate patterns though the same codeword can be shared by different activity feature vectors. Once a codebook is designed, the index numbers of the codewords are used as symbols to apply on the HMM. The index number of the closest codeword from the codebook is used as a symbol to represent a feature vector. Hence, every activity image is assigned a symbol. For instance, K image sequences of T length are converted to K sequences of T symbols. The symbols are the observations O.

B. Activity Training and Recognition

The obtained symbol sequences are used to train HMMs to learn the proper model for each activity such as left hand up-down-HMM or right hand up-down-HMM. We choose four-state left to right HMMs in this study to model sequential events of human activities based on our investigations.

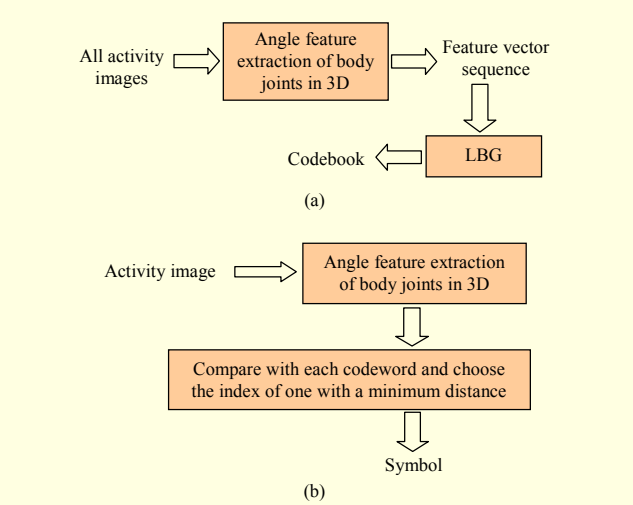

Fig. 11. Basic steps for (a) codebook generation and (b) symbol selection.

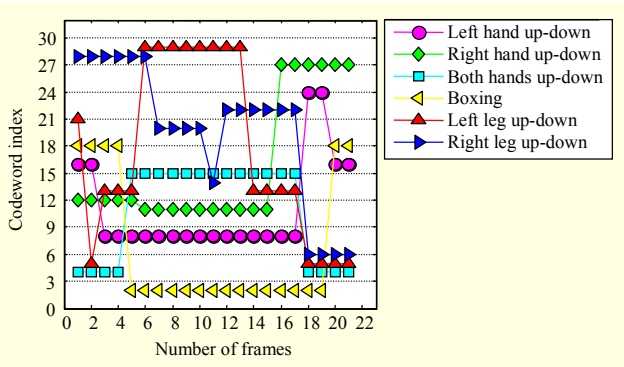

Fig. 12. Patterns of codeword indices for test feature vector sequence of different activities.

For recognition of a 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

$$\text{decision} = \underset{i=1,2,...,M}{\text{argmax}} \{L_i\}, \tag{16}$$

$$L_i = \Pr(O|H_i), \tag{17}$$

where L_i indicates the likelihood of i-th HMM H_i and M number of activities. More details regarding human activity training and testing through HMMs are available in our previous works [4], [5].

III. Experimental Setups

We assume that every training and testing activity video clip contains a single person performing a single activity. As there are no standard stereo camera-based video databases available for HAR, we built our own database for eight different activities (that is, left hand up-down, right hand up-down, both hands up-down, boxing, left leg up-down, right leg up-down, walking, and sitting) to be trained and recognized via the proposed approach. A total of 15 and 40 image sequences of each activity were prepared to use for training and recognition, respectively.

We started our experiments with the traditional binary silhouette-based HAR [1], [3], [4]. After background subtraction, the region of interests containing the binary silhouettes were extracted from every frame. As PCA is generally used for binary silhouette feature extraction, it was applied on all the activity silhouettes for dimension reduction as well as global feature extraction [4]. ICA is superior to PCA by extracting the local features [4], and hence it was also utilized to obtain better performance than PCA. After PCA and ICA, 150 features were considered in the feature space as it is acknowledged that having a greater number of features results in better performance. Finally, the extracted features from the activity image sequences were applied in combination with HMMs for training and recognition. For further experiments, we replaced the binary silhouettes with depth ones and applied the same feature extraction techniques with an HMM to achieve better HAR performance. Thus, to compare the performance of the proposed system with the binary and depth silhouette-based approaches, PC and IC-based experiments were designed in this regard. The same number of training and testing sequences as 3D model-based experiments were applied for both the binary and depth silhouette-based activity recognition.

IV. Experimental Results and Discussions

Since the binary silhouettes from some activities used in our experiments are similar to each other, the recognizer shows the poor results as shown in Table 1.

Usually, ICA is considered to be better than PCA for binary silhouette feature extraction [4], and our binary silhouette-based experiments also demonstrate the superiority of ICA over PCA, showing better HAR performance. Moreover, during the experiments, different activities shared the same binary silhouettes and hence produced ambiguities that results in poor recognition rates in both the cases of ICA and PCA-based approaches.

As depth silhouettes represent a human body better than the binary ones, we continued our experiments to the depth silhouette-based human activity recognition. ICA outperforms PCA regarding the binary silhouette-based HAR [4], [5]; hence, we preferred to apply ICA on the depth silhouettes for

Table 1. Experimental results using binary silhouette features.

Approach	Activity	Recognition rate (%)	Mean (%)	Standard deviation
PCA-based HAR	Left hand up-down	47.50	58.12	19.03
	Right hand up-down	55		
	Both hands up-down	60		
	Boxing	20		
	Left leg up-down	60		
	Right leg up-down	67.50		
	Walking	70		
	Sitting	85		
ICA-based HAR	Left hand up-down	47.50	64.06	18.03
	Right hand up-down	60		
	Both hands up-down	67.50		
	Boxing	30		
	Left leg up-down	72.50		
	Right leg up-down	72.50		
	Walking	75		
	Sitting	87.50		

Table 2. Experimental results using depth silhouette features.

Approach	Activity	Recognition rate (%)	Mean (%)	Standard deviation
PCA-based HAR	Left hand up-down	82.50	80.63	4.38
	Right hand up-down	80		
	Both hands up-down	75		
	Boxing	77.50		
	Left leg up-down	80		
	Right leg up-down	80		
	Walking	80		
	Sitting	90		
ICA-based HAR	Left hand up-down	85	83.44	3.77
	Right hand up-down	82.50		
	Both hands up-down	80		
	Boxing	80		
	Left leg up-down	82.50		
	Right leg up-down	80		
	Walking	87.50		
	Sitting	90		

Table 3. Experimental results using 3D joint-angle features.

Approach	Activity	Recognition rate (%)	Mean (%)	Standard deviation
Joint-angle based HAR	Left hand up-down	87.50	92.81	3.65
	Right hand up-down	97.50		
	Both hands up-down	87.50		
	Boxing	95		
	Left leg up-down	92.50		
	Right leg up-down	95		
	Walking	92.50		
	Sitting	95		
LDA of joint-angle based HAR	Left hand up-down	97.50	98.13	1.77
	Right hand up-down	100		
	Both hands up-down	95		
	Boxing	97.50		
	Left leg up-down	97.50		
	Right leg up-down	100		
	Walking	97.50		
	Sitting	100		

better HAR. Table 2 presents the depth silhouette-based experimental results, showing both the ICA and PCA-based approaches. The experimental results show the superiority of the depth silhouettes over the binary ones and ICA over PCA.

However, we continued our HAR study to the 3D model-based features where far better recognition performance than the binary and depth silhouette-based approaches were obtained as listed in Table 3. We first applied the joint angles directly with HMM and achieved superior recognition rate, 92.81%, over the binary and depth silhouette-based approaches, proving the joint angles are better features for HAR. Since LDA is a powerful tool to find out the underlying feature space to classify the feature vectors linearly, a better feature space can be created applying LDA classification over the 3D model-based joint-angle features. Hence, the LDA-based experiments over the joint-angle features show the highest recognition rate, 98.13%, over others. Thus, our proposed approaches using a 3D model produces much better recognition performance for the complex activities that are not discernable with the binary and depth silhouette-based approaches.

V. Conclusion

In this paper, we have proposed a novel approach for human activity recognition with an HMM utilizing 3D body joint angles directly estimated from a time-series stereo images without optical markers, their inverse kinematic analysis, and multiple cameras. Our proposed HAR system, utilizing the body joint angles as features for HAR, shows superior recognition performance over the conventional approaches of utilizing binary or depth body silhouettes. Our experimental results on the eight different activities reach the mean recognition rate of 98.13%, whereas the conventional binary and depth silhouette-based approaches achieved 64.06% and 83.44% at best. Our marker-free HAR system should be practical in many applications in the fields of smart homes and video-games.

References

[1] J. Yamato, J. Ohya, and K. Ishii, "Recognizing Human Action in Time-Sequential Images using Hidden Markov Model," *IEEE Int. Conf. Comput. Vision Pattern Recognition*, 1992, pp. 379-385.

[2] S. Carlsson and J. Sullivan, "Action Recognition by Shape Matching to Key Frames," *IEEE Comput. Soc. Workshop on Models Versus Exemplars in Comput. Vision*, 2002, pp. 263-270.

[3] F. Niu and M. Abdel-Mottaleb, "View-Invariant Human Activity Recognition Based on Shape and Motion Features," *IEEE 6th Int. Symp. Multimedia Software Eng.*, 2004, pp. 546-556.

[4] M.Z. Uddin, J.J. Lee, and T.-S. Kim, "Independent Shape Component-Based Human Activity Recognition via Hidden Markov Model," *Appl. Intellig.*, vol. 33, no. 2, 2009, pp. 193-206.

[5] M.Z. Uddin et al., "Human Activity Recognition Using Independent Component Features from Depth Images," *5th Int. Conf. Ubiquitous Healthcare*, 2008, pp. 181-183.

[6] T.B. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Comput. Vision and Image Understanding*, vol. 81, no. 3, 2001, pp. 231-268.

[7] M.W. Lee and R. Nevatia, "Human Pose Tracking in Monocular Sequence Using Multilevel Structured Models" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, 2009, pp. 27-38.

[8] C. Chen et al., "3D Human Pose Recovery from Image by Efficient Visual Feature Selection," *Comput. Vision Image Understanding,* vol. 115, no. 3, 2010, pp. 290-299.

[9] I. Chang and S.-Y. Lin, "3D Human Motion Tracking Based on a Progressive Particle Filter," *Pattern Recognition*, vol. 43, no. 10, 2010, pp. 3621-3635.

[10] P.R. Horaud et al., "Human Motion Tracking by Registering an Articulated Surface to 3D Points and Normals," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 1, 2009, pp. 158-164.

[11] L. Sigal L, A.O. Balan, and M.J. Black, "HUMANEVA: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Motion," *Int. J. Comput. Vision*, vol. 87, no. 1-2, 2010, pp. 4-27.

[12] N.D. Thang et al., "Estimation of 3-D Human Body Posture via

Co-Registration of 3-D Human Model and Sequential Stereo Information," *Applied Intell.*, DOI: 10.1007/s10489-009-0209-4, 2010.

[13] J. Cech and R. Sara, "Efficient Sampling of Disparity Space for Fast and Accurate Matching," *IEEE Conf. Comput. Vision Pattern Recognition*, 2007, pp. 1-8.

[14] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vision*, vol. 57, no. 2, 2004, pp. 137-154.

[15] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, 2003, pp. 564-577.

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

[17] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, 2001, pp. 1222-1239.

[18] E.W. Djikstra, "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, vol. 1, 1959, pp. 269-271.

[19] T. Toyoda and O. Hasegawa, "Random Field Model for Integration of Local Information and Global Information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, 2008, pp. 1483-1489.

[20] M.Z. Uddin, J.J. Lee, and T.-S. Kim, "An Enhanced Independent Component-Based Human Facial Expression Recognition from Video," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, 2009, pp. 2216-2224.

[21] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Commun.*, vol. 28, no. 1, 1980, pp. 84-95.

Md. Zia Uddin received his BSc in computer science and engineering from International Islamic University Chittagong, Bangladesh, 2004. He completed his MS leading to a PhD from the Department of Biomedical Engineering, Kyung Hee University, Rep. of Korea, 2011. His research interests include pattern recognition, image processing, computer vision, and machine learning.

Nguyen Duc Thang received his BE in computer engineering from Posts and Telecommunications Institute of Technology, Vietnam. He is currently working toward his MS leading to a PhD in the Department of Computer Engineering at Kyung Hee University, Rep. of Korea. His research interests include artificial intelligence, computer vision, and machine learning.

Jeong Tai Kim received his BSc, MSc, and PhD in architectural engineering from Yonsei University, Korea, in 1977, 1979, and 1985, repectively. Since 1980, he has been a faculty member of the Department of Architectural Engineering, Kyung Hee University, Rep. of Korea. He was the president of the Acoustical Society of Korea in 2005 and the president of the Korea Institute of Ecological Architecture and Environment from 2007 to 2008. He is an associate editor of Indoor and Built Environment and an editorial member of Building and Environment. He has been directing the Center for Sustainable Healthy Buildings as an Engineering Research Center (ERC) supported by National Research Foundation of Korea (NRF) funded by Ministry of Education, Science, and Technology.

Tae-Seong Kim received the BS in biomedical engineering from the University of Southern California (USC) in 1991, MS degrees in biomedical and electrical engineering from USC in 1993 and 1998, respectively, and the PhD in biomedical engineering from USC in 1999. After his postdoctoral work in cognitive sciences at the University of California, Irvine, in 2000, he joined the Alfred E. Mann Institute for Biomedical Engineering and Department of Biomedical Engineering at USC as a research scientist and research assistant professor. In 2004, he moved to Kyung Hee University, Rep. of Korea, where he is currently an associate professor in the Biomedical Engineering Department. His research interests have spanned various areas of biomedical imaging, including magnetic resonance imaging (MRI), functional MRI, E/MEG imaging, DT-MRI, transmission ultrasonic CT, and magnetic resonance electrical impedance imaging. Lately, he has started research work in proactive computing at the u-Lifecare Research Center and the Center for Sustainable Healthy Buildings. Dr. Kim has published more than 60 peer reviewed papers and 120 proceedings. He holds 3 international patents. He is a member of IEEE, KOSOMBE, and Tau Beta Pi and listed in Who's Who in the World '09-'11 and Who's Who in Science and Engineering '11-'12.