Silhouette History and Energy Image Information for Human Movement Recognition

Mohiuddin Ahmad, Irine Parvin and Seong-Whan Lee*
Department of Electrical and Electronic Engineering, Khulna University of Engineering and Technology, Khulna-9203, Bangladesh
*Department of Computer Science and Engineering, Korea University, Seoul 136-713, Republic of Korea
Email: ahmad@eee.kuet.ac.bd

***Abstract*— In this paper, we propose spatio-temporal silhouette representations, called silhouette energy image (SEI) and silhouette history image (SHI) to characterize motion and shape properties for recognition of human movements such as human actions, activities in daily life. The SEI and SHI are constructed by using the silhouette image sequence of an action. The span or difference of the end time and start time is used to make the SHI. For addressing the human shape variability, we used the variation of the anthropometry of the person. We extract the features based on geometric shape moments. We tested our approach successfully in the indoor and outdoor environment. Our experimental results show that the proposed method of human action recognition is robust, flexible and efficient.**

***Index Terms*— Anthropometry, human action recognition, shape variability, silhouette energy image, silhouette history image**

I. Introduction

Recognition of human actions from multiple views by the classification of image sequences has the applications in video surveillance and monitoring, human-computer interactions, model-based compressions, video retrieval in various situations. Typical situations include scenes with moving or clutter backgrounds, stationary or non-stationary camera, scale variation, starting and ending state variation, individual variations in appearance and cloths of people, changes in light and view-point, and so on. These situations make the human action recognition a challenging task. Several human action recognition methods have been proposed in the last few decades. Detailed surveys can be found in [1], [2], [3], [4], where different methodologies of human action recognition, human movement, etc, are discussed. Based on these reviews, researchers either use human body shape information or human body motion information for action recognition. Researchers either used an explicit body shape model or did not use any body shape model, for action recognition. Our approach can be considered as a combination of shape and motion based representation without using any prior body shape model.

This paper is based on "SEI and SHI Representations for Human Movement Recognition," by M. Ahmad and Z. Hossain, which appeared in the Proceedings of the 11th Int'l Conf. on Computer and Information Technology (ICCIT), Khulna, Bangladesh, December 2008. © 2008 IEEE.

The standard approach for human action recognition is to extract a set of features from each image sequence frame and use these features to train classifiers and to perform recognition. Therefore, it is important to consider the appropriateness and robustness of features of action recognition in varying environment. Actually, there is no rigid syntax and well-defined structure for human action recognition available. Moreover, there are several sources of variability that can affect human action recognition, such as variation in speed, viewpoint, size and shape of performer, phase change of action, scaling of persons, and so on. In addition, the motion of the human body is non-rigid in nature. These characteristics make human action recognition a sophisticated task. Considering the above circumstances, we consider some issues that affect the development of models of actions and classifications, which are as follows: (i) An action can be characterized by the local motion of human body parts, (ii) an action can be illustrated by the silhouette image sequence of the human body, which can be regarded as global motion flow, (iii) the trajectory of an action from different viewing directions is different; some of the body parts (part of hand, lower part of leg, part of body, etc) are occluded due to view changes, and (iv) human actions depend on several variability, such as anthropometry, method of performing the action, speed, phase variation (starting and ending time of the action), and camera view variations such as zooming, tilting, and rotating.

Among various features, the motion of the body parts and human body shape play the most significant roles for recognition. Motion based features can represent the approximation of the moving direction of the human body and human action can be effectively characterized by motion rather than other cues, such as color, depth, and spatial features. In the motion-based approach, the motion information of the human such as optic flows, affine variation, filters, gradients, spatial-temporal words, and motion blobs are used for recognizing actions. Motion-based action recognition had been performed by several researchers; a few of them are [7] [15] [16] [19]. The authors in [19] used the local space-time features and integrated these representations with SVM classification schemes for recognition. Ke et al. [12] applied spatio-temporal volumetric feature that efficiently scan video sequences in space and time. Dollár et al. [7] proposed

doi:10.4304/jmm.5.1.12-21

the approach to detect sparse space-time interest points based on separable linear filters for behavior recognition. Niebles et al. [16] used local space time features for unsupervised learning of human actions.

However, motion-based techniques are not always robust in capturing velocity when motions of the actions are similar for the same body parts. On the other hand, the human body silhouette represents the pose of the human body at any instant in time, and a series of body silhouette images can be used to recognize human action correctly, regardless of the speed of movement. Different descriptors of shape information of motion regions such as points, boxes, silhouettes, and blobs are used for recognizing or classifying actions. Several researchers performed action recognition using shapes or silhouettes, such as [5] [6]. Bobick and Davis [5] proposed the motion energy image (MEI) and motion history image (MHI) for human movement representation and recognition and were constructed from the cumulative binary motion images. We propose silhouette energy image (SEI) and silhouette history image (SHI) which gives shape information with global motion information but MEI and MHI give only motion information.

In addition of shape and motion, several variabilities that occurred frequently are also responsible for human action recognition. Sheikh and Shah [20] explicitly identified three sources of variability in action recognition, such as viewpoint, execution rate, and anthropometry of actors and they used the 3D space with thirteen anatomical landmarks for each image. In contrast to their work, we explicitly define and employ the anthropometry variation, camera observations (zooming of a person, slanting body, and rotation of human body), speed variations and multiple views variation of the action. Related works have typically concentrated on the variability in viewpoint [18] by deriving view invariant features or proposing a view invariant algorithm.

We utilize the global shape motion features in addition to some variability's for recognizing the periodic as well as non-periodic actions. The global shape motions are extracted from geometric shape of models. Therefore, based on the combined information of global motion, sources of variability's, and multiple views, human action recognition is more robust and flexible. We propose to recognize several actions of humans in the daily life from multiple views learning of global motion features using the multiclass support vector machine (MCSVM). The actions modeling and classification in this work involve both the Korea university full body gesture database (FBGDB) [8] and the KTH database (KTHDB) [14]. Of particular interest is the detection method, which we use for the recognition of several daily actions of elderly people for human-robot interaction (HRI) or similar applications.

A. Overview of the system

The proposed movement recognition system is shown in Fig. 1. In our system, we assume that silhouette images are correctly extracted and the time span of an action

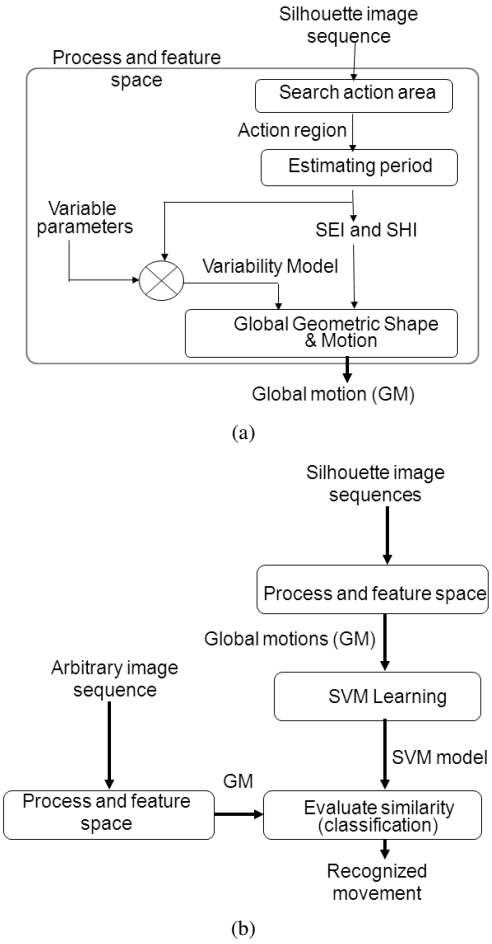


Figure 1. Illustration of the human movement recognition system. (a) Process and features space (b) Complete human movement recognition procedure.

is not greater than 10 seconds of 25 frame rate. For more time, we use correlation technique to delete some frames for generating SEI. From the silhouette images we estimate the temporal boundary (i.e. period or duration) of an action. Depending on the temporal boundary, the SEI and SHI are constructed from the silhouette images. Using the variable parameter(s) and the SEI, variability models are generated. The models are characterized by global shape and motion description. We learn a single action for multiple views global motion descriptors by using a multiclass SVM classifier and generate a unified description of the model for each action. For recognizing actions, we classify (using the similarity of features) descriptions using action models.

B. Organization of the paper

This paper is organized as follows: Section II presents action representation using SEI and SHI and variability generation in our system. Section III discusses global motion descriptors of SEI and SHI models. Section IV shows the classification approach of human actions. Section V presents experimental results and discussions of the selected approaches. Finally, conclusions are drawn in Section VI.

II. SEI, SHI, AND HUMAN SHAPE VARIABILITY

The silhouette history image represents the shape history information of an image sequence. This history information represents the time history for performing an action. SEI and SHI gives the idea of motion and action recognition by watching silhouettes.

A. Human action or Human movement

Human action is the movement of human body parts for performing a task within a short period of time. The action may be simple or complex depending on the number of body limbs involved in the action. Many actions performed by humans have cyclic nature and they show periodicity of short duration. Besides, many actions show single occurrence or non-periodic with time frame of specific length (i.e. duration). We consider human actions daily performed which are almost cyclic in nature, either multiple cyclic (period=nT) or periodic actions (different types of walking, running, jogging, etc), and single occurrence or non-periodic actions (bowing, raising the hand, sitting on the floor, etc). Under the above circumstances, it is possible to transform a human action in the spatio-temporal space or 3D space, into a 2D spatial space, where the 2D space contains spatial information with temporal information.

B. Silhouette energy image (SEI)

We define the silhouette energy image as the image where the time variation of an action is represented by the body shape information. Let us assume $x_t(x,y) = f(x,y,t)$ is the silhouette image in a sequence at time t, which includes an action under duration or a period as shown in Fig. 2(a). Therefore, SEI (SEI =$S(x,y)$) is defined by (1). Moreover, the standard deviation image as well as motion variation expressions are given by (2) and (3).

$$S(x,y) = \frac{1}{t_e - t_s}\int_{t_s}^{t_e} x_t dt = \frac{1}{F}\sum_{t=1}^{F} x_t(x,y) \quad (1)$$

$$\sigma(x,y) = \sqrt{\frac{\int_{t=t_s}^{t_e} x_t^2 dt}{t_e - t_s} - \Big(\frac{1}{t_e - t_s}\int_{t=t_s}^{t_e} x_t dt\Big)^2} \quad (2)$$

$$C_v(x,y) = \frac{\sigma(x,y)}{S(x,y)} \quad (3)$$

Here, t_s and t_e are the starting and ending states of an action. Alternately, we can represent total number of frame by F and hence the period is F. Therefore, the period or duration becomes $F = t_e - t_s$. Since the average 2D image stores the global motion distribution and orientation of the silhouette images, we can designate this as a SEI. The number of frames in the action depends on the person, time, and type of action. Since, we use the average of the time sequence silhouette images; the normalized variation affects are very low. Fig. 2(a) shows the sample silhouette images with the SEI of the "raising hand" action along with the variation of motion. The silhouette energy image (SEI), standard deviation and motion variation images are shown in Fig. 2(b), Fig. 2(c) and Fig. 2(d) for different span of times, respectively. This representation shows the shape as well as motion changes of an action. The SEI represents an action model (AT), due to the following reason: (1) The energy of a pixel at every point is a result of an action formation; (2) Each silhouette represents the unit energy of a human action at any instance; (3) It determines the energy distribution of an action.

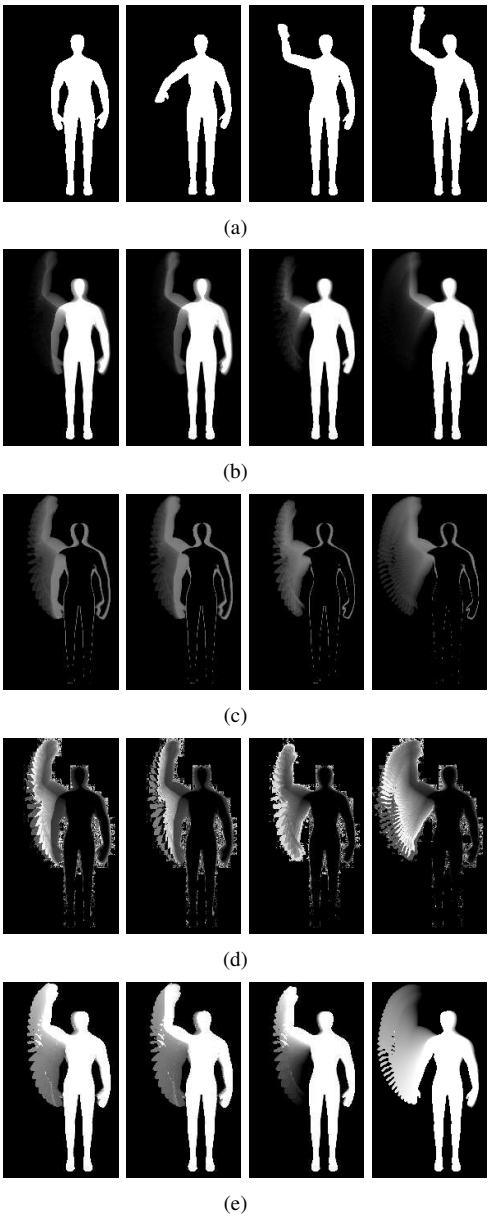


Figure 2. Human action representation using Silhouette energy image (SEI) and silhouette history image (SHI) (a) Some key frames of an action (b) SEI at different time span (50 frames, 60 frames, 20 frames, and 50 frames) (c) Standard deviation images at the same time span of SEI. (d) Motion variation images at same time span of SEI. (e) SHI at the same time span of SEI.

C. Silhouette history image (SHI)

Silhouette history image (SHI) refers to the shape variation of the image sequence in time. By SHI, we represent how the silhouette of an image sequence is moving. It does not show only the motion image representation as [5] but also it represents the global motion orientation of an action at any instant of time. The shape of the person at present state as well as the person's global motion orientation is visualized by the SHI.

From the image sequences of same action, we show the SHI in Fig. 2(e). We construct these representations by using plastic model of human. From both SHI, we can show the action orientation or simply the direction of the action. Apart from the MHI representation in [5], SHI represents both body shape and global motion change. The brighter region represents the recency of the global human body shape. The current human body pose is related to the previous one, so we use the method of MHI [5]. We use silhouette images instead of motion images for making a SHI given by (4).

$$H_t(x,y,t)=\begin{cases}\tau, & \text{if } x_t(x,y)>1\\ max(0,H_t(x,y,t-1)), & \text{otherwise}\end{cases} \tag{4}$$

where τ is the duration of temporal extension to previous silhouette image and $x_t(x,y)$ is a silhouette image indicating the region of human at time t. SEI and SHI are regarded as the action models (ATs) in our approach.

D. Justification of Representation

We represent human action by silhouette image sequence, called silhouette energy image (SEI) which saves both space and computation time for recognition of actions. We mentioned that the normalized variation affects are very low, that means SEI is less sensitive to noise in individual frame in a sequence.

Since silhouette image may contain noise, therefore, we can consider that silhouette image, $x_t(x,y)$ is a combination of a original image, $o_t(x,y)$ and a noise image, $n_t(x,y)$. We also consider that at every pair of coordinates (x,y) the noise at different moments t is identically distributed and has no correlation. We consider that total number of frame is needed for performing an action is F.

Now, the silhouette image is given by

$$x_t(x,y)=o_t(x,y)+n_t(x,y) \tag{5}$$

Moreover, we consider that $n_t(x,y)$ satisfies the following distribution:

$$n_t(x,y)=\begin{cases}n_{1t}(x,y):\begin{cases}P\{n_t(x,y)=-1\}=p\\ P\{n_t(x,y)=0\}=1\text{–}p,\\ \quad\text{if } o_t(x,y)=1\end{cases}\\ n_{2t}(x,y):\begin{cases}P\{n_t(x,y)=1\}=p\\ P\{n_t(x,y)=0\}=1\text{–}p,\\ \quad\text{if } o_t(x,y)=0\end{cases}\end{cases} \tag{6}$$

We have

$$E\{n_t(x,y)\}=\begin{cases}-p, & \text{if } o_t(x,y)=1\\ p, & \text{if } o_t(x,y)=0\end{cases} \tag{7}$$

and

$$\sigma^2_{n_t(x,y)}=\sigma^2_{n_{1t}(x,y)}=\sigma^2_{n_{2t}(x,y)}=p(1-p) \tag{8}$$

We consider that SEI is constructed from F frames where $o_t(x,y)=1$ at pixel (x,y) only in G frames. We have

$$\begin{aligned}S(x,y)&=\frac{1}{F}\sum_{t=1}^{F}x_t(x,y)\\ &=\frac{1}{F}\sum_{t=1}^{F}o_t(x,y)+\frac{1}{F}\sum_{t=1}^{F}n_t(x,y)\\ &=\frac{G}{F}+\bar{n}(x,y)\end{aligned} \tag{9}$$

Now, the noise in silhouette image is

$$\bar{n}(x,y)=\frac{1}{F}[\sum_{t=1}^{G}n_{1t}(x,y)+\sum_{t=G+1}^{F}n_{2t}(x,y)] \tag{10}$$

Therefore, the first moment of the noise is

$$\begin{aligned}E\{\bar{n}(x,y)\}&=\frac{1}{F}[\sum_{t=1}^{G}E\{n_{1t}(x,y)\}+\sum_{t=G+1}^{F}E\{n_{2t}(x,y)\}]\\ &=\frac{1}{F}[G(-p)+(F-G)p]\\ &=\frac{(F-2G)p}{F}\end{aligned} \tag{11}$$

The second moment of the noise (i.e. variance) is

$$\sigma^2_{\bar{n}_t(x,y)}=E\{[\bar{n}(x,y)-E\{\bar{n}(x,y)\}]^2\} \tag{12}$$

$$\begin{aligned}&=\frac{1}{F^2}E\{[\sum_{t=1}^{G}n_{1t}(x,y)-E\{n_{1t}(x,y)\}]\\ &+\sum_{t=G+1}^{F}[n_{2t}(x,y)-E\{n_{2t}(x,y)\}]^2\}\\ &=\frac{1}{F^2}[G\sigma^2_{n_{1t}(x,y)}+(F-G)\sigma^2_{n_{2t}(x,y)}]\\ &=\frac{1}{F^2}\sigma^2_{n_t(x,y)}\end{aligned} \tag{13}$$

Therefore, the mean of the noise in SEI varies between $-p$ and p depending on G while its variability (second moment of noise) decreases. If $G=F$ at (x,y) (all $o_t(x,y)=1$), then $E\{\bar{n}_t(x,y)\}=-p$. If $G=0$ at (x,y) (all $o_t(x,y)=0$), then $E\{\bar{n}_t(x,y)\}=p$. At (x,y), the mean of the noise in SEI is the same as that in individual silhouette image, but the noise variance reduces so that the probability of outliers is reduced. If $G=0\sim N$ at (x,y), $E\{\bar{n}_t(x,y)\}=p\sim -p$. Therefore both the mean and the variance of noise in SEI are reduced compared to individual silhouette image at these locations. At the extreme, the noise in SEI has zero mean and reduced variance where $G=F/2$. As a result, SEI is less sensitive to silhouette noise in individual frames, i.e. the normalized variation affects are low.

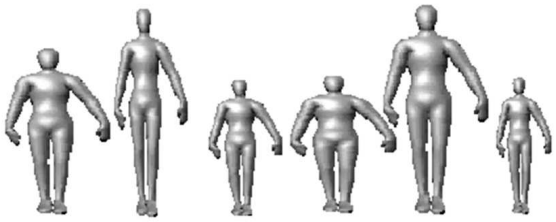


Figure 3. Anthropometric variation images with different body width and height.

E. Variability of body shape

To consider the diversity of modeling (learning) and classifying actions, we consider multiple human body shape, and the produced models are known variability models (VTs). Usually, anthropometry variation follows no specific rule. We have approximated the *variability of anthropometry* to different actions. Figure 3 shows the example of anthropometry variation. Due to different girth[1] and height variations, human action models should adapt anthropometry. Therefore, the anthropometry variation of an action represents the width and height variation of a person. Due to these variations, human action recognition should adapt anthropometry. We define basic eight[2] sets of anthropometric variations, $S(x,y)|_a$ by (14).

$$S(x,y)|_a = \begin{cases} S(x-a,y), & \text{Results Hg} \\ S(x,y-b), & \text{Results Hh} \\ S(x+a,y), & \text{Results Lg} \\ S(x,y+b), & \text{Results Lh} \\ S(x-a,y+b), & \text{Results HgLh} \\ S(x+a,y-b), & \text{Results LgHh} \\ S(x-a,y-b), & \text{Results HgHh} \\ S(x+a,y+b), & \text{Results LgLh} \end{cases} \quad (14)$$

where, H=Higher, g=girth, L=Lower, and h=height. Moreover, a & b are width and height variation parameters with $a \approx 0.067R \sim 0.1R$ and $b \approx 0.067C \sim 0.1C$. Given the original SEI template of size $R \times C$. The variability models can be constructed by resizing the human body using bilinear interpolation method. As an example, Results HgLh represents the generation of higher girth lower height from the original action model.

An action can be performed at a different speed or frame rates, which is the number of silhouette images in a sequence. We model action by considering two factors for different frame rate. These include (i) change of the number of frames and (ii) pixel variations. Following these factors, the *speed variability* images can be modeled by (15). The expression does not rigorously follow the speed variation, but it approximates the variation of speed of an action.

$$S(x,y)|_s = \begin{cases} S(x,y)(\frac{F}{F+n})\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\pi\sigma^2}} \\ S(x,y)(\frac{F}{F-n})\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\pi\sigma^2}} \end{cases} \quad (15)$$

where n is a small time unit and $n \ll F$. Here, F is the total frame required for performing an action. The first expression is used when we model an action by greater frame number (frame$>$ F) and the later expression is valid for less frame number (frame$<$ F).

At the time of performing an action, the *view observation* (such as the position, orientation, scaling of the persons) and *view variations* can be changed. Therefore, we can consider three kinds of camera parameters variation and they include (1) distance from camera - it refers to the varying scale of the persons body position from camera, (2) tilting motion or slanting motion - human body may in slanting position when a human performs an action, (3) human body rotation - body rotation during the action. The parameters (1) and (2) can be modeled by using affine transforms. The parameter (3) variation is modeled by projection geometry. Suppose, a point $\mathbf{x} = (x,y)$ in the coordinate system of shape is affine transformed to a point $\mathbf{x}_a = (x_a, y_a)$ in the imaging plane's coordinate system, then variability models $S(x,y)|_c = S(x_a, y_b)$ from the camera observations are given by (16).

$$S(x,y)|_c = s\Big(\begin{bmatrix} d+s_x & s_y-r \\ r+s_y & d-s_x \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}\Big) \quad (16)$$

where, d, r, t_x, t_y, s_x, and s_y represent the dilation (scaling or divergence), rotation, translation along x-axis, translation along y-axis, shear component along-x, and shear-component along-y, respectively. By modeling the coefficient parameters, the diverse representation of view-observation models can be achieved. Furthermore, we consider that an action can be seen from multiple views. In this case, addition of features from multiple views for the same action can be a solution.

III. Global Geometric Shape and Motion

We express the global shape motion description by the model orientation, span and elevation of motions, geometric and orthogonal moments. Therefore, we express the geometric shape motions by $\{s_g, s_z, s_d\}$. The symbols are defined in the following subsections.

A. Geometric moments

Moments and function of moments have been utilized as pattern feature in pattern recognition applications. Such features capture global information about the image and do not require close boundaries as required by Fourier descriptors. By positioning the center of mass (COM), we can differentiate the motions for each action. Since this feature should be independent of the location of a person, then we consider the relative appropriate position for each action. Hu [10] introduced seven nonlinear functions, h_i, where $i = 1, 2, \cdots, 7$ defined on regular moments using

[1]Girth is the band or strap that encircles the body of a human or animal to fasten something (as a saddle) on its back. It is used in 3D analysis. The width is considered to be the projection of girth.

[2]Theoretically, a huge numbers of anthropometric variations can be created by using the anthropometry variation parameter.

central moments that are translation, scale, and rotation invariant. To achieve the consistent camera observations, we used non-orthogonal features, namely Hu moments, $s_g = \{h_1, h_2, \cdots, h_7\}$ [10] which are slightly modified to extract specific characteristics. The values of h_i are extracted from different combinations of η_{pq} [10], where $p+q$ gives the order of moment function. The value of η_{pq} is shown in (20).

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} = \frac{1}{\mu_{00}^{\gamma}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x-\bar{x})^p (y-\bar{y})^q S(x,y) \quad (17)$$

where,

$$\mu_{00} = m_{00} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} S(x,y) \quad (18)$$

$$(\bar{x}, \bar{y}) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right) \quad (19)$$

$$\gamma = \frac{p+q+2}{2} \text{ and } \forall (p+q) \geq 2 \quad (20)$$

The advantage of a moment's methods is that they are mathematically concise and for the intensity image of action models, they reflect both shape and global motion distribution within it.

B. Zernike moments

The geometric moment shows highly inaccurate results when the image is noisy. Zernike polynomials provide very useful moment kernels, present native rotational invariance and are far more robust to noise. Scale and translation invariance can be implemented using moment normalization. The magnitude of Zernike moments has been treated as shape features because they are rotation invariant. The two-dimensional Zernike moments of an image intensity function $f(\rho, \theta)$ with order n and repetition m is expressed as follows [13].

$$Z_{nm} = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 R_{nm}(\rho) e^{-jm\theta} S(\rho,\theta) \rho d\rho d\theta \quad (21)$$

and

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{(n-|m|)}{2}} \frac{(-1)^s (n-s)! \rho^{n-2s}}{(s)!((n+|m|)/2-s)!((n-|m|)/2)!} \quad (22)$$

where,

$$\rho = \frac{\sqrt{(2x-N+1)^2 + (N-1-2y)^2}}{N} \quad (23)$$

$$\theta = \tan^{-1} \left(\frac{N-1-2y}{2x-N+1} \right) \quad (24)$$

and $0 \leq \rho \leq 1$. For each action model and one given value, we obtain the Zernike moments value. We use the absolute Zernike moment, which is given in (25).

$$Z_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} S(x,y) R_{nm}(\rho) \exp(-jm\theta) \quad (25)$$

as the global motion features, s_z for SEI, SHI, and variability models.

C. Direction of actions

The 2D orientation (direction of major axis, or minor axis) of the motion distribution for every action is different. Thus the relative differences in magnitude of the eigenvalues are an indication of the elongation of the image (SEI or SHI). The global motion orientation is obtained from the eigenvalue λ_i, of the covariance matrix of SEI, SHI and/or variability models.

The covariance matrix is

$$M_c = \begin{pmatrix} \acute{\mu}_{20} & \acute{\mu}_{11} \\ \acute{\mu}_{11} & \acute{\mu}_{02} \end{pmatrix} \quad (26)$$

The eigenvalue of the covariance matrix is given by

$$\lambda_i = \frac{\acute{\mu}_{20} + \acute{\mu}_{02}}{2} \pm \frac{\sqrt{4\acute{\mu}_{11}^2 + (\acute{\mu}_{20} - \acute{\mu}_{02})^2}}{2} \quad (27)$$

where,

$$\acute{\mu}_{pq} = \mu_{pq} / \mu_{00} \quad (28)$$

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x-\bar{x})^p (y-\bar{y})^q S(x,y) \quad (29)$$

We have considered the projection of major and minor axis orientation and the direction of the major axis $s_d = \{proj.(\lambda_i)\}$ as global features for SEI, SHI, and variability models.

IV. CLASSIFICATION OF ACTIONS

Human action recognition can be considered as a pattern classification problem that can be solved by measuring the similarity between training features and testing features. The classification can be carried out by different process, namely, normal Bayes classifier, k-nearest neighbor classifier, and support vector machine (SVM) classifier derived from feature vectors. Of these, SVM has high generalization capabilities in many tasks, especially in terms of object recognition. To model and classify actions, we used multi-class SVM classifiers. Each action consists of multiple views and multiple scenarios motion descriptors.

Consider the pattern recognition problem of training samples (x_1, y_1), (x_2, y_2), $\cdots$, (x_l, y_l), where $\mathbf{x}_i$, $i = 1, 2, \cdots, l$ is a vector and $y_i \in \{1, 2 \cdots, k\}$ represents the class of samples. The multi-class support vector machines (SVM) [21] require the solution of the following optimization problem: minimize

$$\phi(\omega, \xi) = \frac{1}{2} \sum_{m=1}^{k} \omega_m \cdot \omega_m + C \sum_{i=1}^{l} \sum_{m \neq y} \xi_i^m \quad (30)$$

with constraints

$$\begin{gathered} (\omega_{y_i} \cdot x_i) + b_{y_i} \geq (\omega_m \cdot x_i) + b_m + 2 - \xi_i^m, \\ \xi_i^m \geq 0,\ i = \{1, 2, \cdots, l\},\ m \in \{1, \cdots, k\} \setminus y_i \end{gathered} \quad (31)$$

where C is the penalty parameter, l is the number of training data, k is the number of classes, y_i is the class of the ith training data ω points perpendicular to the separating hyperplane, b is the offset parameter to increase the margin, and ξ is the degree of misclassification of the datum x_i. This gives the decision function:

$$f(x) = arg \max_{m=1,\cdots,k} [\omega_m \cdot x + b_m] \quad (32)$$

This optimization problem is solved by finding the saddle point of the Lagrangian [21], [17]. In applying the multi-class SVM, the motion descriptors of SEI and variability models are classified into the defined classes. The learning and classification part consists of a training module and classification module for global motion descriptor. The training data of the motion descriptors of the models are divided into defined classes manually. The MCSVM predicts the class label information for arbitrary action. Before applying MCSVM, we normalize the global motion descriptor.

V. Experiment Results and Discussion

A. Databases

The FBGDB [8] contains 14 representative full body actions in the daily life of 20 performers. In the database, all the performers are elderly persons (both male and female) with ages ranging from 60 to 80. The database contains 3D motion data and 2D data. The 2D data consists of both video data and 2D silhouette data. As an example, the sample images for left view are shown in Fig. 4.

The KTHDB is one of the largest databases with sequences of human actions taken over different scenarios [14]. The database contains six types of human actions, performed several times by 25 subjects in four scenarios: outdoors (s1), outdoors with scale variation (s2), outdoor with different cloths (s3), and indoor (s4). The database contains 2391 sequences. The sample images are shown in Fig. 5. The image sequences have the spatial resolution of 160×120 pixels and have a length of four seconds in average.

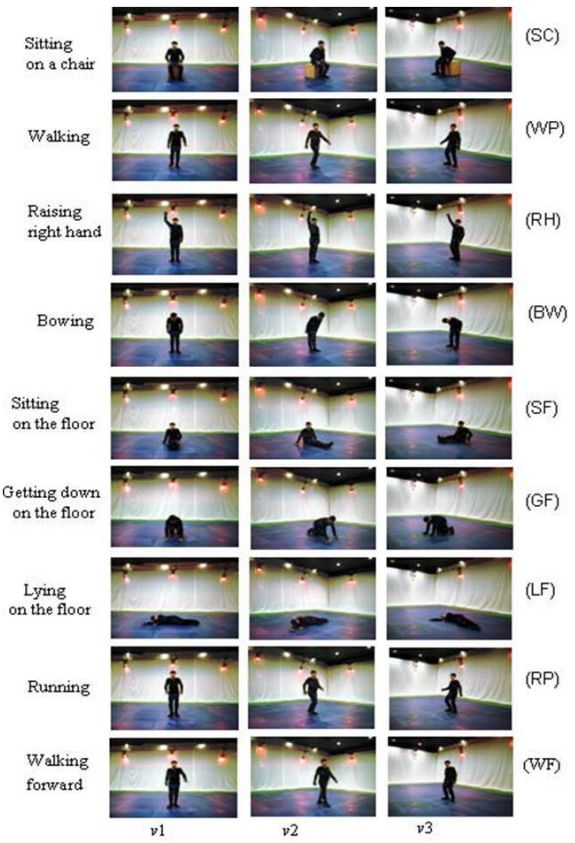


Figure 4. Korea University Gesture database (FBGDB) on all specified actions in three different views.

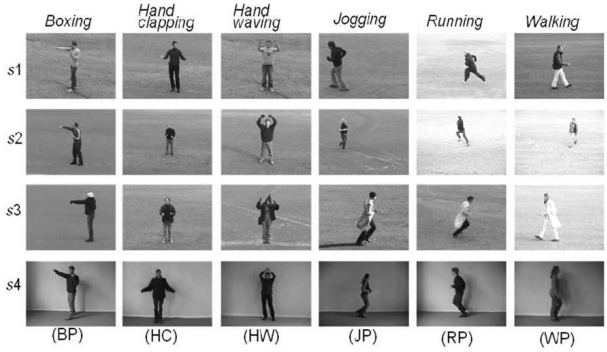


Figure 5. Royal Institute of Technology (KTH) human action database on all specified actions in 4 different scenarios.

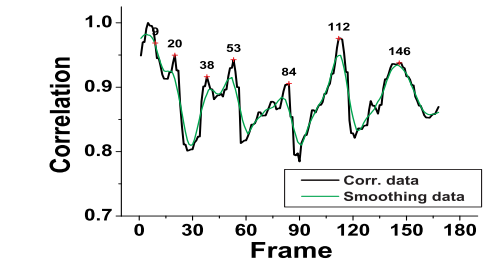
B. Estimation of duration of an action

We estimate the period or duration by correlation or using the variation in pixel distribution in the silhouette image sequences. Let us consider that p is the period. Therefore, the periodicity relationship becomes, $f(t + p) = f(t)$, where $f(t)$ is the motion of a point, or energy of an image at any time t. A non-periodic function is one that has no such period, instead we use the duration of action.

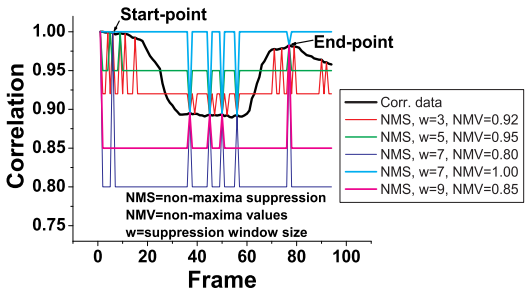
The brief algorithm for detecting period (or duration) is as follows: First, assume reference frame is the 1st $\sim$ 5th frame of the given silhouette image sequence. Second, find the similarity (i.e. cross-correlation or energy) of silhouettes. Third, apply smoothing operation to the similarity plot for periodic action and extract peak points. For non-periodic action, we apply non-maxima suppression (NMS) method and make decision to extract the peak points (starting point and ending point). We choose multi-scale non-maxima window size (w) for selecting the peak points, where non-maxima values (NMV) are chosen arbitrarily. Now, the period is given by the difference between starting point and ending point as illustrated in Fig. 6.

C. Examples of SEI

We consider 9 actions from the FBGDB, where the actions are key actions occurring in daily life. The typical action models are shown in Fig. 7. The brighter parts indicate more silhouette energy and the less bright parts indicate less silhouette energy of the action. From the



(a)



(b)

Figure 6. Periodicity (or duration) detection from silhouette image sequences (FGBDB). (a) Running with multiple cycles (t_s, t_e)={(20, 53), (53, 84), (84, 112), (112,146)} with smoothing. (b) Raising the right hand in a single occurrence ($t_s = 6$, $t_e = 77$).

TABLE I.
CRR OF EACH ACTION AND VIEW OF FGBDB

View	SC	WP	RH	BW	SF	GF	LF	RF	WF
v1	.98	.57	.86	1.0	.71	.57	1.0	.57	.86
v2	.85	.44	1.0	.86	.86	.86	1.0	.86	.86
v3	.97	.86	1.0	1.0	.86	1.0	1.0	.57	.72
vA	.94	.57	.94	.99	.76	.76	.90	.76	.76

action models, the motion distribution of each action is clearly understood.

Fig. 8 shows typical action models and corresponding significant motion variation over the models of KTHDB. For each action, the motion variation and shape is different. The motion variation clarifies the actions clearly.

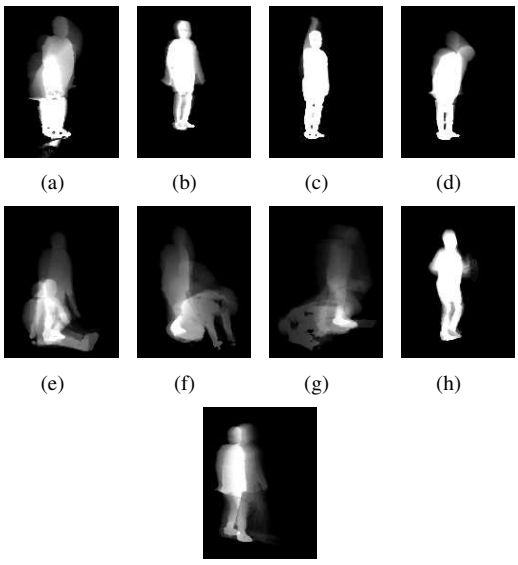
D. Classification results

We define the accuracy or correct recognition rate (CRR) by (33). The expression for CRR can be written as

$$CRR = \frac{N_c}{N_a} \times 100(\text{in percentage}) \qquad (33)$$

Where, N_c is the total number of correct recognition sequences while N_a is the number of total action sequences.

Table I shows the action recognition results of FBGDB using MCSVM classifiers where we use the global motions for each view. We use 9 subjects, 9 actions, and 4 views variation for testing (vA represents arbitrary view). As can be seen, there is a clear separation among different



(a) (b) (c) (d)

(e) (f) (g) (h)

(i)

Figure 7. Human action model of the specified actions for the FGBDB. (a) Sitting on a chair (SC). (b) Walking at a place (WP). (c) Raising the right hand (RH). (d) Bowing (BW). (e) Sitting on the floor (SF). (f) Getting down on the floor (GF). (g) Lying down on the floor (LF). (h) Running at a place (RP). (i) Walking forward (WF).

TABLE II.
CRR OF EACH ACTION AND VIEW OF KTHDB

Scen.	BP	HC	HW	JP	RP	WP
s1	.98	.97	.97	.91	.74	.82
s2	.98	.88	.94	.83	.61	.78
s3	1.0	.96	.97	.74	.81	.81
s4	.95	.94	.93	.81	.67	.80

kinds of actions. The overall CRRs of v1, v2, v3, and vA are 79.34, 84.12, 89.47 and 81.53 respectively, of FBGDB. We use SEI, SHI, and variability models to evaluate the performance.

We also have tested our approach by using the KTHDB, since it is one of the largest human action databases and several researchers used this database. We have tested 8 subjects, 6 actions, and 4 scenarios and each scenario contains 2 or 3 action sequences. Table II shows the recognition of each action in various scenarios for global shape motions. The CRRs of s1, s2, s3, s4, sA are 89.33, 83.17, 87.50, 88.3, 87.50 respectively for KTHDB, where sA is arbitrary scenario.

We use 21 image sequences for classification of each action and each view in case of FBGDB. For arbitrary view recognition, we use more than 60 sequences for each action. It is important to mention that in some cases, the motion of the elderly persons is similar. In our method, it is shown that by using the 2D action model with variability selection, the action recognition is more robust, since we use the natural actions of humans, with emphasis on elderly persons (FBGDB). The movement of elderly person's is significantly different than that of young people. For example, the speed and style of walking and running of elderly people are very similar.

We test the system performance without generating

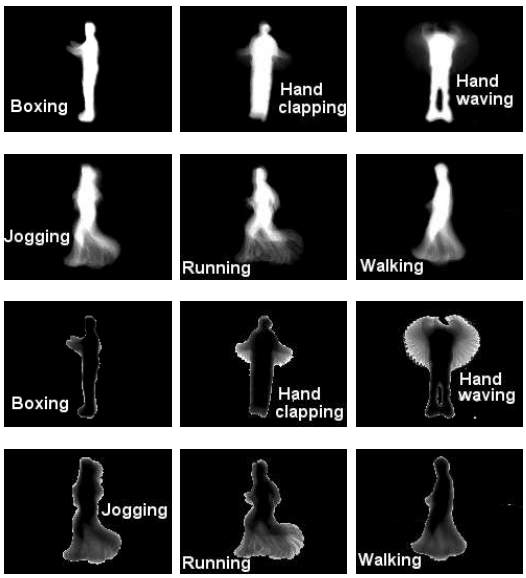


Figure 8. Human action models and corresponding motion distribution (KTHDB). First and second rows show the action models. Third and fourth rows show the motion distribution of corresponding actions.

TABLE III.
COMPARISON OF ACTION RECOGNITION

Method	**Recognition accuracy**	**Scenarios**
Niebles et al. [16]	81.50	all scenarios
Dollár et al. [7]	81.17	all scenarios
Jiang et al. [11]	84.43	all scenarios
Schüldt et al. [19]	71.72	all scenarios
Ke et al. [12]	62.96	all scenarios
Our method	87.50	all scenarios

adaptable models, and we make a comparison of performance among SEI (AT), variability models (VT), and the combined models (AAT). As an example, the performance (in CRR) of AT, VT, and AATs are 80%, 84.33%, and 89.33% respectively. The performance of AAT is significantly better than AT.

We compare our works with some state-of-art action recognition approaches by using the same database and similar test sequences but different methods. For example, we compare our method with [7], [11], [12], [16], and [19] using KTHDB. Our results by global shape motions flow are compared with their results by spatio-temporal filters, volumetric features, spatio-temporal words, and local space time features. The overall comparison of different methods is listed in Table III. Compared to the mentioned researches, our approach yields the best recognition results.

VI. CONCLUSION

We proposed a novel method for recognizing human action using the silhouette energy image and silhouette history image with the variation of anthropometric models. The SEI represented the energy content of the silhouette images of an action and SHI represented the energy history of the silhouettes of that action. The anthropometry variation provided a more natural and robust environment for human action recognition, using an advanced human-machine interface. Moreover, by adapting the shape variation, incomplete actions and partial occluded actions were recognized successfully. We recognized human actions in individual and arbitrary views. With global motion features, the action recognition became sparse and flexible and it can be adapted to practical applications of human movement, human action recognition, and so on. Our future work will include the precise detection and recognition of action in complicated situations.

ACKNOWLEDGMENT

This research is partly supported by CASR, Khulna University of Engineering and Technology, Khulna-9203, Bangladesh

REFERENCES

[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A Survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90-126, October 2006.
[2] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems, Man, and Cybernetics-Part C: Applications and Review*, vol. 34, No. 3, pp. 334-352, August 2004.
[3] D. Gavrila, "The visual analysis of human movement: a survey," *Computer Vision and Image Understanding*, vol. 73 no. 1, pp. 82-98, 1999.
[4] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81. no. 3, pp. 231-268, 2001.
[5] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.
[6] S. Carlsson and J. Sullivan, "Action Recognition by Shape Matching to Key Frames," *Proc. IEEE Workshop on Models versus Exemplars in Computer Vision*, pp. 263-270, 2002.
[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-temporal Filters," *Proc. IEEE Int'l Workshop VS-PETS*, pp. 65-72, 2005.
[8] The KU Gesture Database *http://gesturedb.korea.ac.kr/.*
[9] C. -W. Hsu and C. -J. Lin, "A Comparison Methods for Multiclass Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415-425, March 2002.
[10] M-K. Hu., "Visual Pattern Recognition by Moment Invariants," *IRE Trans. On Information Theory*, IT-8, pp. 179-187, 1962.
[11] H. Jiang, M. S. Drew, and Z. N. Li, "Successive Convex Matching for Action Detection," *Proc. Computer Vision and Pattern Recognition*, vol. 2, pp. 1646-1653, 2006.
[12] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. Int'l Conf. Computer Vision*, pp. 166-173, 2005.
[13] A. Khotanzad and Y. H. Hong, "Invariant Image Recognition by Zernike Moments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489-497, May 1990.
[14] The KTH Database, *http://www.nada.kth.se/cvap/actions/.*
[15] O. Masoud and N. Papanikolopoulos, "Recognizing Human Activities," *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, pp. 157-162, July 2003.

[16] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Proc. Proc. BMVC*, vol. 3, pp. 1249-1258, September 2006.
[17] C. J. Lin and R. C. Weng, "Simple Probabilistic Predictions for Support Vector Regression," Technical Report, National Taiwan University, 2004.
[18] V. Parameswaran and R. Chellappa, "View Invariants for Human Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 613-619, 2003.
[19] C. Schüldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. IEEE Conf. ICPR*, vol. 3, pp. 32-36, 2004.
[20] Y. Sheikh and M. Shah, "Exploring the Space of a Human Action," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 144-149, October 2005.
[21] J. Westons and C. Wtkins, "Support Vector Machines for Multiclass Pattern Recognition," *Proc. 7th European Symposium on Artificial Neural Networks*, pp. 219-224, 1999.

Mohiuddin Ahmad received his BS degree with Honors in Electrical and Electronic Engineering from CUET, Bangladesh and his MS degree in Electronics and Information Science from Kyoto Institute of Technology of Japan in 1994 and 2001, respectively. He received his PhD degree in Computer Science and Engineering from Korea University in 2008. From August 1995 to October 1998, he served as a lecturer in the Department of Electrical & Electronic Engineering at KUET, Bangladesh. In June 2001, he joined the same Department as an Assistant Professor and he has been serving in the same department. His research interests include human action modeling and recognition, biomedical engineering, image and signal processing and their applications in computer vision and the pattern recognition related fields.

Irine Parvin received her BS degree in Electrical and Electronic Engineering from KUET, Bangladesh in 2004. Now she is a Masters Degree candidate in the same Department. Her research interests include computer vision and pattern recognition, image and signal processing, etc.

Seong-Whan Lee received his BS degree in Computer Science and Statistics from Seoul National University, Seoul, Korea, in 1984, and his MS 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 biometrics.