Image and Vision Computing 28 (2010) 814–824

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Variable silhouette energy image representations for recognizing human actions ☆

Mohiuddin Ahmad [a], Seong-Whan Lee [b,*]

[a] Department of Electrical and Electronic Engineering, Khulna University of Engineering and Technology, Khulna 9203, Bangladesh
[b] Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-713, Republic of Korea

ARTICLE INFO

Article history:
Received 27 February 2009
Received in revised form 30 June 2009
Accepted 18 September 2009

Keywords:
Silhouette energy image
Action recognition
Variability action models
Daily life actions
Global motion description

ABSTRACT

Recognizing human actions is an important topic in the computer vision community. One of the challenges of recognizing human actions is describing for the variability that arises when arbitrary view camera captures human performing actions. In this paper, we propose a spatio-temporal silhouette representation, called silhouette energy image (SEI), and multiple variability action models, to characterize motion and shape properties for automatic recognition of human actions in daily life. To address the variability in the recognition of human actions, several parameters, such as anthropometry of the person, speed of the action, phase (starting and ending state of an action), camera observations (distance from camera, slanting motion, and rotation of human body), and view variations are proposed. We construct the variability (or adaptable) models based on SEI and the proposed parameters. Global motion descriptors express the spatio-temporal properties of combined energy templates (SEI and variability action models). Our construction of the optimal model for each action and view is based on the support vectors of global motion descriptions of action models. We recognize different daily human actions of different styles successfully in the indoor and outdoor environment. Our experimental results show that the proposed method of human action recognition is robust, flexible and efficient.

1. 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 several papers including [1].

We consider the approach for recognizing actions 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, view-point, 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 role 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 and depth. In the motion-based approach, the motion information of the human such as optic flows, affine variation, filters, gradients,

☆ A preliminary version of the paper has been presented in the IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, September 2008.

* Corresponding author. Tel.: +82 2 3290 3197; fax: +82 2 926 2168.
E-mail addresses: ahmad@eee.kuet.ac.bd (M. Ahmad), swlee@image.korea.ac.kr (S.-W. Lee).

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.09.018

spatial-temporal words, and motion blobs are used for recognizing actions. Motion-based action recognition has been performed by several researchers; a few of them are [4,12,13,15]. 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 [2,3]. Bobick and Davis [2] 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. Han [20] proposed the gait energy image for individual recognition. Another gait recognition, called motion silhouette image (MSI) was proposed by Lam in [21]. We propose silhouette energy image (SEI) from the silhouette image sequence for human action recognition.

Besides motion and body shape, several variability that occur frequently are also responsible for human action recognition. Sheikh and Shah [16] explicitly identified three sources of variability in action recognition, such as view-point, execution rate, and anthropometry of actors and they used the 3D space with thirteen anatomical landmarks for each image. Related works have typically concentrated on the variability in view-point [14] by deriving view invariant features or proposing a view invariant algorithm.

During the action recognition of persons, we utilize the 2D information of global shape motion features in addition to several variability's for recognizing the periodic as well as non-periodic or single occurrence actions. The global shape motions are extracted from geometric shape of models. Therefore, based on the combined information of global motion, sources of variability, 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 multi-class support vector machine (MCSVM).

1.1. Motivation

Our work is motivated by the ability of humans to utilize periodic and non-periodic motions to perform several actions, which are frequently used in the daily life. It is well recognized that many actions are periodic in nature. This periodic nature of human actions can be analyzed using the shape of human beings, since body parts, as well as shapes, can change while performing particular actions. Shape analysis plays an important role in action recognition, gait recognition, etc. In many situations, we are interested in the movement of human body silhouette (shape) over time. The shape changing of humans describes the nature of human's motion and shows the action or activity performed by humans. This change of shape over time is considered as a result of global motion of the shape and deformations. We consider this global motion change by compact representation where we accumulate all time information into static time information, i.e. 2D information. This static time information of the resulting image provides an important cue for global and local motion characteristics, such as motion distribution, motion orientation, shape deformation, etc. By using appropriate variable parameters, we consider more relational characteristics of each action.

1.2. Human action and variability

We define human action as the movement of humans for performing a task within a short period of time. In this paper, we represent a human action by silhouette energy images (SEI) which is constructed from a sequence of silhouette image. We consider that similar prototype actions, called variability templates (VTs) or models are generated from SEI and variability parameters. The variability parameters are: (1) anthropometry, (2) variation of phase (varying starting and ending state of actions), (3) speed variation of an action, and (4) camera observations (zooming of the person, slanting motion, and body rotation). During the representation of SEI, we utilize the period or duration of an action. The period or duration represents the difference between starting and ending states of an action. The period of an action may be represented by half cycle, one cycle or several cycles. The period of an action might be a half cycle, when the remaining half cycle predictably follows the same pattern of the previous half cycle. Strictly speaking, both representations of actions are not the same, but we can consider them as approximately similar. We also considered multiple view variations (action recognition from multiple viewing directions). Moreover, the person's clothing, occlusion, etc. affect the recognition. All the above mentioned factors are closely related with human action representation and recognition.

1.3. Motion description: global

A SEI is constructed by using the sequence of silhouette images. Therefore, in case of SEI, we can consider the energy of a pixel is a local motion descriptor. So, an average of a set of data (energy) can present a more informative characteristic of the local image contents. On the contrary, a global descriptor can also be defined as a number (or a value) which characterizes the whole image content. The 'semi-global descriptor' refers to the use of many values as opposed to one single value, such as a set of values in multiscale window centered on the concerned point. Therefore, multiple values can characterize an image with more information for comparing image sets. We use global descriptor of SEI and corresponding VTs for characterizing and classifying human actions.

1.4. Contributions

In this paper, we use the image similarity to recognize human actions. The major contributions are as follows:

1. We use the 2D representation of human action model, called SEI, accumulating time varying silhouette images at a unit time for action recognition. Therefore, action representation using SEI saves both storage space and time.
2. We introduce the explicit variability action model, for considering different forms of the same action, for human action recognition. Four important factors are considered, which include anthropometry of persons, speed of an action, the starting and ending phase of an action, and the camera observations (zoom, scale, and rotation). Moreover, multiple view variations are adapted, which make the human action recognition more robust.
3. Typical scenarios with homogeneous-stationary camera, scale variations, appearance and cloths variations, multiple views, and incomplete actions are recognized.

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.

1.5. Overview of the system

In our system, we assume that silhouettes of an image sequence are correctly captured. From the silhouette image sequence, we estimate the temporal boundary (i.e. period or duration) of each

action [19]. Depending on the temporal boundary, an action model (i.e. SEI) is constructed by the silhouette image sequence. Using the variability parameters and the SEI, VTs are generated. The models are characterized by global motions. We learn an action for multiple view global motion descriptors by using a MCSVM, and generate SVM models for specified actions. For recognizing actions, we classify (using the similarity of features) descriptions using SVM models. The actions modeling and classification in this work involve both the Korea university full body gesture database (FBGDB) [5] and the KTH database (KTHDB) [11]. Our proposed action recognition approach is shown in Fig. 1.

1.6. Organization of the paper

This paper is organized as follows: Section 2 presents the action representation of our approach. Section 3 represents the generation of variability models in our system. Section 4 discusses global motion descriptors of combined models. Section 5 shows the classification approach of human actions. Section 6 presents experimental results and discussions of the selected approaches. Finally, conclusions are drawn in Section 7.

2. Silhouette energy image (SEI) representation

Human action is the movement of humans 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 have considered 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 (duration = p) 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 temporal information. Let us assume $x_t = f(x, y, t)$ is the silhouette image in a sequence at time t, which includes an action under a duration or a period. Therefore, SEI (SEI $= s(x, y)$) is defined by Eq. (1): moreover, the standard deviation image as well as motion variation expressions are given by Eqs. (2) and (3) which gives motion information and variation.

$$s(x,y) = \frac{1}{p}\sum_{t_s}^{t_e} x_t, \begin{cases} t_e - t_s = nT \\ t_e - t_s = p \end{cases} \tag{1}$$

$$\sigma(x,y) = \sqrt{\frac{\sum_{t=t_s}^{t_e} x_t^2}{p} - \left(\frac{1}{p}\sum_{t=t_s}^{t_e} x_t\right)^2} \tag{2}$$

$$c_v(x,y) = \frac{\sigma(x,y)}{s(x,y)} \tag{3}$$

Here, t_s and t_e are the starting and ending states of an action. 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 effects are very low. Fig. 2 shows the sample silhouette images with the SEI of the "rais-

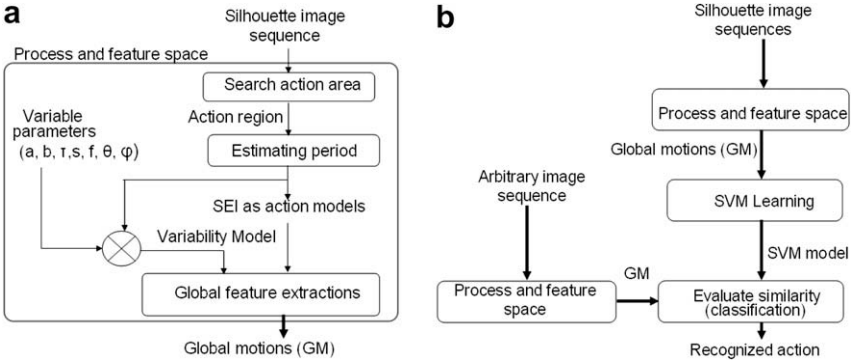

Fig. 1. Illustration of the human action recognition system. (a) Process and features space. The variability models are generated from original action model by controlling each variable or adaptable parameter. (b) Complete human movement recognition procedure.

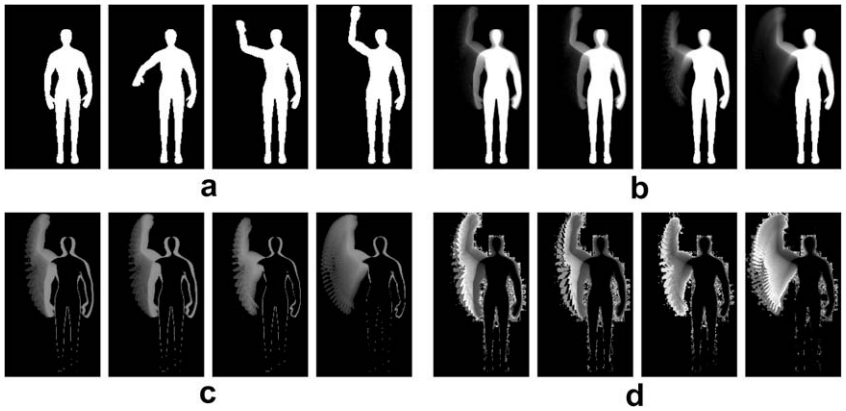

Fig. 2. Human action representation using silhouette energy image (SEI). (a) Some key frames of an action (b) SEI at different time span; (i) 1–51 (50 frames), (ii) 1–60 (60 frames), (iii) 20–40 (20 frames), and (iv) 60–110 (50 frames). (c) Standard deviation images at the same time span of SEI. (d) Motion variation images at same time span of SEI.

ing hand" action along with the variation of motion. 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.

3. Variability action models

3.1. What are variability action models?

The variability action models or variability templates (VTs) are defined as noise action models or complementary action models that are generated by using SEI and variability parameters. If the representation of an action derived from different variability or adaptability parameters (anthropometry, execution rate, phase, camera observation) are similar, then this representation is said to be robust for adaptability of these parameters. The original action model is not a unique representation for an action, since several sources of variability affect human action recognition, such as size and shape of performer, phase change of action, execution rate variation, clothing of the performers, scale variation, camera observation, and view-point variation, etc.

3.2. Types of variability

To consider the diversity of modeling (learning) and classifying actions, we consider multiple variability models or templates (VTs) or complementary action models.

3.2.1. Anthropometry variability

In general, human actions are performed irrespective of the shape (appearance) of the performer. Usually, anthropometry variation follows no specific rule. We have approximated the adaptability of anthropometry to different actions. Fig. 3 shows the example of anthropometry variation. Due to different girth[1] and height variations, human action models should adapt anthropometry. These variations are modeled using anthropometric variation. Due to the variation of human anthropometry as shown in Fig. 3, we can define eight[2] sets of anthropometric variations. Mathematically, we can express these variations by using sub-matrices, or super-matrices, or the combination of sub and super-matrices that are resized into original size for getting the anthropometric variability images. Each resize is done by bilinear interpolation method. Let the SEI is represented by $A[0:R-1,0:C-1]$, where R and C are the number of row and column of the SEI, respectively. In our case, the sub-matrix becomes $B[a:R-a-1,b:C-b-1]$ and the super-matrix becomes $B[0-a:R+a-1,0-b:C+b-1]$.

$$s(x,y)|_a = \begin{cases} B[a:R-a-1,0:C-1], & \text{Results Hg} \\ B[0:R-1,b:C-b-1], & \text{Results Hh} \\ B[0-a:R+a-1,0:C-1], & \text{Results Lg} \\ B[0:R-1,0-b:C+b-1], & \text{Results Lh} \\ B[a:R-a-1,0-b:C+b-1], & \text{Results HgLh} \\ B[0-a:R+a-1,b:C-b-1], & \text{Results LgHh} \\ B[a:R-a-1,b:C-b-1], & \text{Results HgHh} \\ B[0-a:R+a-1,0-b:C+b-1], & \text{Results LgLh} \end{cases} \tag{4}$$

[1] Girth is the band that encircles the body of a human or animal to fasten something 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 model can be created using the anthropometric parameters.

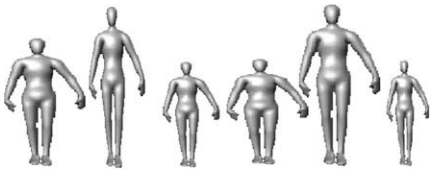

Fig. 3. Anthropometry variation images with different body width and height.

where a and b are the anthropometric variation parameters, which represents how many rows and columns we will cut or add to generate the anthropometric variability models. In (4), H, L, g, and h means higher, lower, girth, and height respectively.

3.2.2. Speed variability

An action can be performed at a different speed or using a different number of frames, which are the number of shape images in an input sequence. By considering temporal transformation, we can adopt the action at a different speed. In the case where speed or execution rate of human actions vary, we can consider two phenomena:

1. The action can be performed at a speed faster or slower than the standard speed, i.e. number of frames. Let us consider the person's velocity is s and N frames are needed to perform the action, then, without loss of generality, we can say, the execution rate of an action is inversely proportional to the speed of that action, i.e. $s \propto \frac{1}{N}$. Therefore, a linear relationship exists between the number of frames and an action.
2. Every pixel in the SEI shows motion variation due to the performed action. So, due to execution rate changes, the motion at each pixel never changes linearly, because any added frame to the sequence is not linear to the previous and next frames. For simplicity, we can model this variation using Gaussian function.

Suppose an action is performed by more than two persons. The execution time of the action depends on the actor's performance. Based on condition 1 (condition 2 is NULL), we can say, $N_1 s_1 = N_2 s_2$, where N_1, N_2 are the required time (period) for performing the action using speed s_1 and s_2, respectively. If N is the typical time for the action, then the relationship between N_1 ($N_1 > N$) and N is given by $N_1 = N - n$. In a similar sense, actor-2 performs the same action with the period N_2 and the relationship between N_2 and N is given as $N_2 = N + n$. Let us assume that, n is a small time unit, where $N \gg n$. Now, according to condition 2 (condition 1 is NULL), due to the nonlinear relationship, we introduce a small variation of motion in the pixel using the Gaussian kernel function in the spatial space of the silhouette image. A 3×3 spatial Gaussian kernel is used. Therefore, using the two above assumptions, we can model the variable or adaptable speed of the action by using the following Eq. (5). This does not rigorously follow the speed variation, but it approximates the variation of speed of action.

$$s(x,y)|_s = \begin{cases} s(x,y)\left(\frac{N}{N+n}\right)\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\pi\sigma^2}}, & \text{Action period} > N \\ s(x,y)\left(\frac{N}{N-n}\right)\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\pi\sigma^2}}, & \text{Action period} < N \end{cases} \tag{5}$$

where n is a small time unit and $n \ll N$ and N is the required time for performing an action.

3.2.3. Phase variability

The variable 'phase variation' refers to an action occurred at different starting and ending state. The starting and ending phase of

an action depends on persons, time, style, and so on. For example, in the 'bowing' action, a person bends the waist at different angles from the reference position, i.e. from a standing position. Therefore, we can express the phase variability models at starting (ϕ_s) and ending (ϕ_e) by Eq. (6).

$$s(x,y)|_\tau = \begin{cases} \frac{1}{p-\phi_s}\sum_{t_s+\phi_s}^{t_e} x_t, & \phi_s \text{ varies} \\ \frac{1}{p-\phi_e}\sum_{t_s}^{t_e-\phi_e} x_t, & \phi_e \text{ varies} \end{cases} \quad (6)$$

In this definition, the parameters ϕ_s and ϕ_e represent the starting and ending phase variation from start and end. Due to phase variation, the starting and ending state of an action changes because of few frames blank (which we can consider incomplete actions). An illustrating situation of phase variation is shown in Fig. 4.

3.2.4. Camera observation and view variations

At the time of performing an action, the position, orientation, scaling of the persons, and view-points can be changed. Therefore, we have considered 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. Besides camera parameter observation, we consider that an action can be seen from several views. Fig. 5a and b illustrate the camera observations and multiple views variation of an action.

The parameters (1) and (2) are modeled by using affine transforms. The parameter (3) variation is modeled by projection geometry. We use affine transformation to simulate a planar shape that undergoes 2D rotation, translation, and scaling. 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 $S(x,y)|_c = S(x_a, y_a)$ from the camera observations are given by Eq. (7).

$$s(x,y)|_c = s\left(\begin{bmatrix} d+s_x & s_y - r \\ r+s_y & d-s_x \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}\right) \quad (7)$$

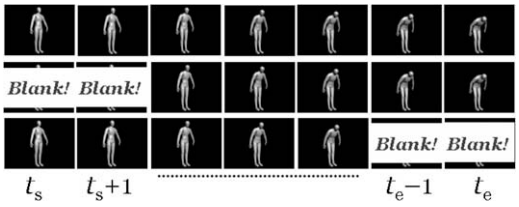

Fig. 4. Phase variation. Top row: complete action. Middle and bottom row: incomplete action. "*Blank!*" refers to some frames missing at start or end. Here, t_s and t_e are the starting and ending state of the action.

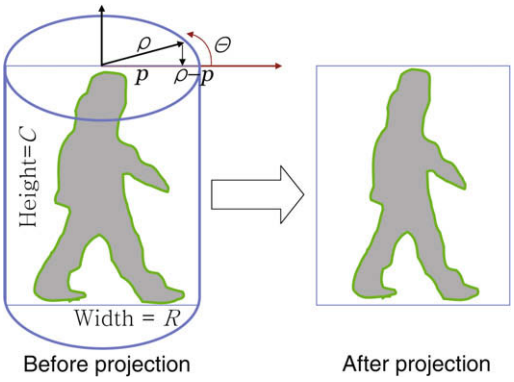

Fig. 6. Illustration of 2D projection geometry.

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. In order to model the rotation of human body, we consider that the width of an image, (R) is the diameter of a cylinder, 2ρ, where ρ is the radius of the cylinder. We also consider that the center line of the image is the center line of the cylinder so that the image can rotate along its axis. The situation is shown in Fig. 6. We get 2D projection image from a 2D image and assume the input image rotates around its y-axis. The rotation angle is $\pm 10^0$. The 2D image after the projection is

$$f_{c.rot}(x,y,t) = f(x + 2\rho(1-\cos\theta), y, t) \quad (8)$$

After resizing the image into the original image size, we get the rotation variations models. By modeling the coefficient parameters, diverse representation and learning of actions can be achieved.

4. Global shape motion descriptions

We define the combined variability templates or models (CVT) as the combination of action models (SEIs or ATs) and variability models (VTs). We described the geometric shape motions by $\{s_g, s_z, v_x, v_y, v_h, s_{k_p}, s_o\}$. The notations are defined in the following subsections.

4.1. 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. Hu [7] introduced seven nonlinear functions, h_i, where $i = 1, 2, \ldots, 7$ defined on regu-

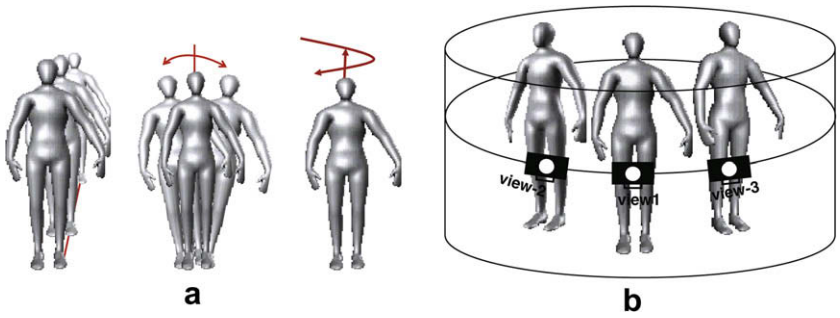

Fig. 5. Observation and view variation of human actions. (a) Camera observations. Left: person's distance from camera (scaling of a person). Middle: slanting position of human body. Right: human body rotation around upward axis. (b) Multiple views variation.

lar moments using central moments that are translation, scale, and rotation invariant. We use $s_g = \{h_1, h_2, h_3, h_4, h_5, h_6, h_7\}$ as the geometric moment features. The advantage of a moment's methods is that they are mathematically concise and for the intensity image of action model, they reflect not only the shape but also the global motion distribution within it.

4.2. 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. The magnitude of Zernike moments has been treated as global motion descriptors because they are rotation invariant. The 2D Zernike moment of the image intensity function $S(\rho, \theta)$ with order n and repetition m is given in [10].

$$Z_{nm} = \frac{n+1}{\lambda_N} \int_0^{2\pi} \int_0^1 R_{nm}(\rho) e^{-jm\theta} s(\rho, \theta) \rho d\rho d\theta \tag{9}$$

where ρ = length of vector from origin, θ = angle between vector ρ and x-axis in ccw direction, $0 \leqslant \rho \leqslant 1$, λ_N is a normalization factor, and $R_{nm}(\rho)$ is radial polynomial. To achieve scale and translation uniformity, the image function $f(x,y)$ is transformed into $g(x,y)$, where $g(x,y) = f\left(\frac{x}{a} + \bar{x}, \frac{y}{a} + \bar{y}\right)$ [10] and hence the Zernike moment be $\widehat{Z}_{nm}$ with $(\bar{x}, \bar{y})$ being the centroid (COM) of $f(x,y)$ and a is a predetermined value. The absolute value of Zernike moment is given by Eq. (10).

$$Z_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} s(x,y) R_{nm}(\rho) \exp(-jm\theta) \tag{10}$$

The value of $\widehat{Z}_{nm}$ is given by Eq. (11). This is obtained by replacing $s(x,y)$ by $g(x,y)$ of Eq. (10).

$$\widehat{Z}_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} g(x,y) R_{nm}(\rho) \exp(-jm\theta) \tag{11}$$

We use $s_z = \{\widehat{Z}_{11}, \widehat{Z}_{10}, \widehat{Z}_{01}, \widehat{Z}_{20}, \widehat{Z}_{02}, \widehat{Z}_{22}, \widehat{Z}_{31}\}$ as Zernike moment features.

4.3. Shape motion distribution

We mentioned that the action model describes the global motion over the duration. For any action, the mean absolute deviation v_x and v_y of the pixels relative to the COM is used for motion description.

$$\{v_x, v_y\} = \frac{\sum_{(x,y) \in s(x,y) \geqslant Th} (x - \bar{x}, y - \bar{y}) s(x,y)}{\sum_{(x,y) \in s(x,y) \geqslant Th} s(x,y)} \tag{12}$$

where Th represents the threshold value. With this motion description, we can distinguish between actions where more body parts are involved in motion (e.g. sitting on floor, getting down on the floor, lying down on the floor, etc.), and an action concentrated in a smaller area where only small parts of the body move (e.g. sitting on a chair, bowing, etc.). Another important feature for describing global motion is the mean intensity of motion, v_h of the pixel distribution. This represents the average absolute height or elevation of motion distribution and is expressed as Eq. (13).

$$v_h = \frac{\sum_{(x,y) \in s(x,y) \geqslant 0} s(x,y)}{\sum_{(x,y) \in s(x,y) \geqslant 0} \max s(x,y) = 255} \tag{13}$$

A large value of v_h indicates very intense motion of the human body parts and a small value indicates minimal motion.

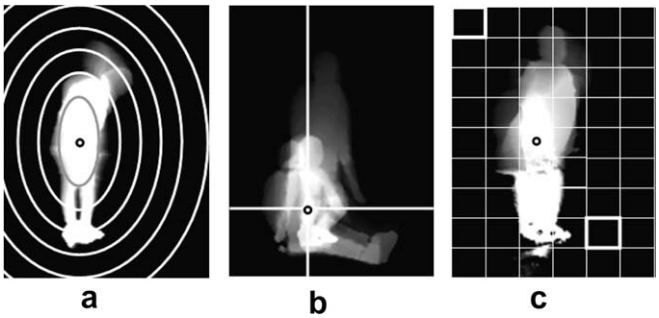

Fig. 7. Multi-geometry partial global motion distribution. (a) Ellipse. (b) Quadrants. (c) Square-block.

4.4. Semi-global motion distribution

It should be noted that the global motion distribution from the center of motion is different for each action, i.e. it results in a different semi-global motion distribution for different kinds of geometry. Each semi-partial distribution (layer) has its own characteristics. According to human body shape, either elliptical, quadrant, or block configuration of motion distribution may be suitable, which is shown in Fig. 7. For any kind of geometry, the motion over successive regions are given by the Eq. (14).

$$s_{k_p} = \frac{1}{n \in B(k_p)} \sum_{(x,y) \in B(k_p)} s(x,y) \tag{14}$$

where $k_p = 1, 2, \ldots, B$ and B is the number of blocks or quadrants depends on geometry p, n is the number of pixels in a block, and k_p represents ellipse (k_e), quadrant (k_q), or block (k_b). For unique representation of motion, we consider three kinds of partial motion distribution, $s_{k_p} = \{s_{k_e}, s_{k_q}, s_{k_b}\}$.

4.5. Shape motion orientation

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). The global motion orientation is obtained from the eigenvalue λ_i, of the covariance matrix of SEI and/or variability models.

The covariance matrix is

$$M_c = \begin{pmatrix} \acute{\mu}_{20} & \acute{\mu}_{11} \\ \acute{\mu}_{11} & \acute{\mu}_{02} \end{pmatrix} \tag{15}$$

The eigenvalue of the covariance matrix is given by

$$\lambda_i = \frac{\acute{\mu}_{20} + \acute{\mu}_{02}}{2} \pm \frac{\sqrt{4\acute{\mu}_{11}^2 + (\acute{\mu}_{20} - \acute{\mu}_{02})^2}}{2} \tag{16}$$

where

$$\acute{\mu}_{pq} = \mu_{pq} / \mu_{00} \tag{17}$$

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^p (y - \bar{y})^q s(x,y) \tag{18}$$

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

5. Classification of actions

The classification of human action can be carried out by different process, namely, Bayes classifier, k-nearest neighbor (kNN) classifier, and support vector machine (SVM) classifier, etc. derived

from feature vectors. Among them, SVM has high generalization capabilities in many tasks, especially in terms of object recognition. SVM is based on the idea of hyperplane classifier that achieves classification by a separating surface (linear or nonlinear) in the input space of data set. SVMs are modeled as optimization problems with quadratic objective functions and constraints. This optimization problem is solved by finding the saddle point of the Lagrangian [18,17]. In applying the Multi-class SVM (MCSVM) [6], the motion descriptors of action models (ATs) and variability models (VTs) are classified into the defined classes. The learning and classification part consists of a training module and classification module for global motion descriptors. The training data of the motion descriptors of the models are divided into defined classes manually. The MCSVM predicts the class label information for an arbitrary action. We employ the one-against-one approach [6] for multi-class classification of actions. Before applying MCSVM, we normalized each motion descriptor. Radial basis function kernel is used here.

6. Experimental results and discussion

6.1. Databases

6.1.1. FBGDB

The FBGDB [5] 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 consists of 2D video data and silhouette data taken at three views: Front view ($v1$), left-side or -45° view ($v2$), and right-side or $+45^\circ$ view ($v3$). The sample images are shown in Fig. 8, where the symbols represent the actions' name.

6.1.2. KTHDB

The KTHDB is one of the largest databases with sequences of human actions taken over different scenarios [11]. 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 sample images are shown in Fig. 9, where the symbols represent the actions' name.

6.2. Period or duration estimation

We estimate the period or duration (i.e temporal boundary) by correlation or using the variation in pixel distribution in the silhouette image sequences, which was discussed in [19]. We can also estimate them manually. Let 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. In case of image, $f(t)\approx f(t+p)$.

The brief algorithm for detecting period (or duration) is as follows: First, assume reference frame is the 1st–10th frame of the given silhouette image sequence. Second, find the similarity (i.e. cross-correlation) of the reference frame to other silhouettes. Alternately, we can also find the energy of each silhouette image. Third, apply smoothing operation to the similarity plot for periodic action and extract peak points. For non-periodic action, we apply

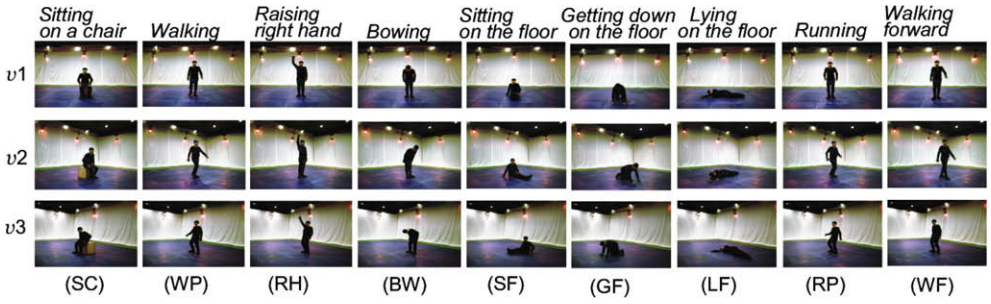

Fig. 8. Example images of FBGDB in three views.

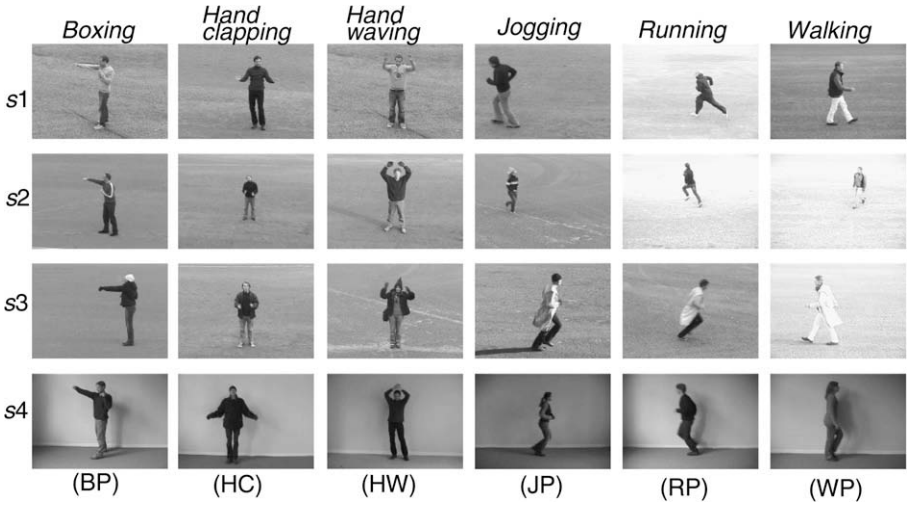

Fig. 9. Example images of KTHDB in four scenarios.

non-maxima suppression (NMS) method and make decision to extract the peak points (starting point and ending point). We choose multiscale 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. 10. The similar explanations were discussed in Ref. [19].

6.3. Action models or SEI

We consider nine actions from FBGDB [5], where the performers are elderly persons and the actions are key actions occurring in daily life. The typical action models (SEI) are shown in Fig. 11. The brighter parts indicate more silhouette energy and the less bright parts indicate less silhouette energy of the action. Fig. 12 shows typical SEIs and corresponding motion variation over the models of KTHDB. For each action, the motion variation and shape is different. The motion variation clarifies the actions clearly.

6.4. Variability parameters

We use the following set of variability parameters and the values for generating the variability models. They are shown in the Table 1.

6.5. Variability models

Theoretically, by changing the variability parameters, we can generate huge number of variability models from SEI. Moreover, similar kind of VTs maintains a high correlation among them. Therefore, the selected numbers of VTs are sufficient for learning actions. Fig. 13 shows typical VTs of a SEI. The number (i)–(viii) means the images from left to right for each sub-figure. The original SEI is shown in Fig. 11e with the period of the action, p which refers to the time for performing an action, N i.e. $p = N$. Fig. 13 briefly described as follows: (a) Anthropometric variational image for adaptation. (i) Hg, (ii) Hh, (iii) Lg, (iv) Lh, (v) HgLh, (vi) LgHh,

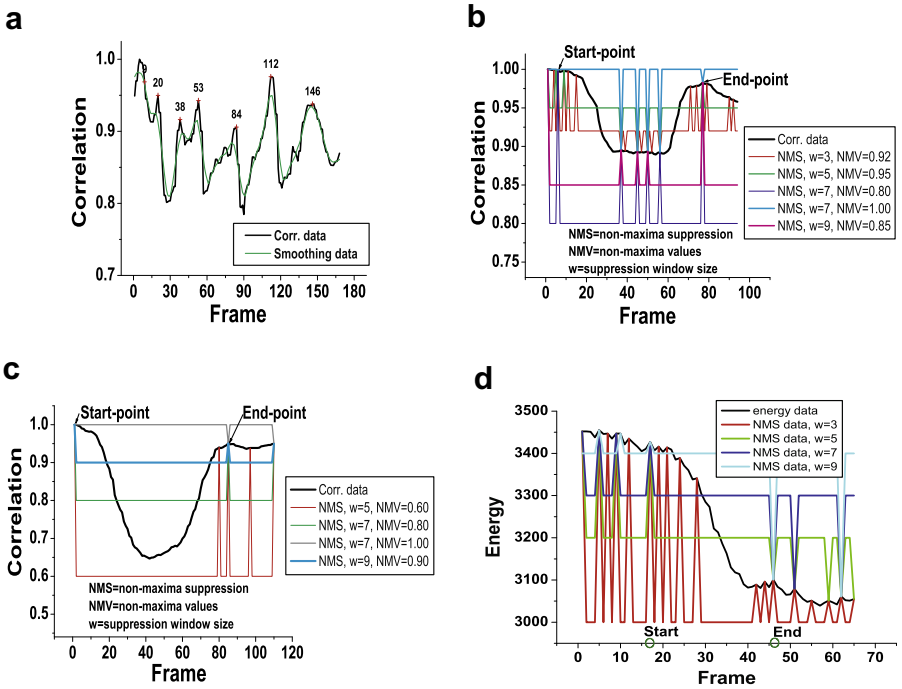

Fig. 10. Periodicity (or duration) detection from silhouette image sequences (KUGDB). (a) Running with multiple cycles $(t_s, t_e) = \{(20, 53), (53, 84), (84, 112), (112, 146)\}$ with smoothing. (b) Raising the right-hand in single occurrence ($t_s = 6,\ t_e = 77$). (c) Bowing in single occurrence ($t_s = 1,\ t_e = 85$). (d) Sitting on a chair in single occurrence ($t_s = 17,\ t_e = 46$) or ($t_s = 9,\ t_e = 50$). This is done by energy calculation of the silhouettes.

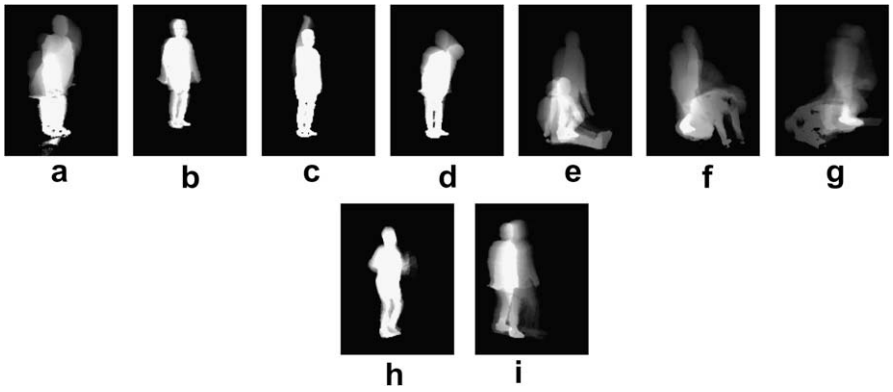

Fig. 11. SEI of the specified actions for the FBGDB. (a) Sitting on a chair (SC). (b) Walking at a place (WP). (c) Raising the right-hand (RH). (d) Bowing (BW). (e) Sitting on the floor (SF). (f) Getting down on the floor (GF). (g) Lying down on the floor (LF). (h) Running at a place (RP). (i) Walking forward (WF).

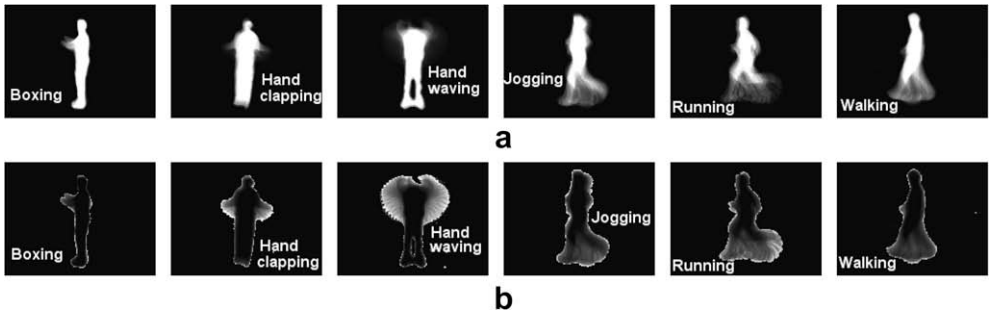

Fig. 12. SEI and corresponding motion distribution (KTHDB). (a) Action models or SEI. (b) Motion distribution of corresponding actions.

Table 1
Variability parameters set.

Variability	Symbols	Typical values
Anthropometry	$\triangle a$, $\triangle b$	$\pm 10\% h$
Speed	s	$\pm 20\% p$
Phase	τ	$15\% p$
Zoom	$\triangle f$	$\pm 20\% f$
Slant	$\triangle \theta$	10^0
Rotation	$\triangle \phi$	10^0

Legend: w = Width, h = height, p = period or duration, f = focal length.

(vii) HgHh, and (viii) LgLh. The variables are described in Section 3.2. (b) Speed variability images. (i) $p = N - 20\%N$. (ii) $p = N - 15\%N$. (iii) $p = N - 10\%N$. (iv) $p = N - 5\%N$. (v) $p = N + 5\%N$. (vi) $p = N + 10\%N$. (vii) $p = N + 15\%N$. (viii) $p = N + 20\%N$. (c) Starting (ϕ_s) and ending (ϕ_e) phase variation models. Here, τ represents a percentage of frames of the period, which refers to the "*Blank!*" frames of an action as stated in Fig. 4. (i) $\phi_s = \tau$. (ii) $\phi_s = 2\tau$. (iii) $\phi_s = 3\tau$. (iv) $\phi_s = 4\tau$. (v) $\phi_e = \tau$. (vi) $\phi_e = 2\tau$. (vii) $\phi_e = 3\tau$. (viii) $\phi_e = 4\tau$. (d) Variable models for zooming a camera. This is an alternative representation of person's distance to the camera. Person's position is varied with the frontward or backward the camera. For original model, assume person's distance from camera, $d = f$. (i) $d = f - 4k$, where k is assumed a specific distance. (ii) $d = f - 3k$. (iii) $d = f - 2k$. (iv) $d = f - k$. (v) $d = f + k$. (vi) $d = f + 2k$. (vii) $d = f + 3k$. (viii) $d = f + 4k$. (e) Variation models of person's slanting position. Maximum slanting angle is $\Delta\theta$ from the vertical position to leftward (Lw) or rightward (Rw) position. The original model has no slanting angle. (i) $Lw = \Delta\theta$. (ii) $Lw = \Delta\theta/2$. (iii) $Lw = \Delta\theta/3$. (iv) $Lw = \Delta\theta/4$. (v) $Rw = \Delta\theta/4$. (vi) $Rw = \Delta\theta/3$. (vii) $Rw = \Delta\theta/2$. (viii) $Rw = \Delta\theta$. (f) Rotation of the person around upward axis. We consider a reasonable limit of horizontal rotation during performing the action. The original model has no rotation. (i) Rotation angle, $\theta = \Delta\theta/k$, where $\Delta\theta$ is the maximum range of rotation and k is the number of images. (ii) $\theta = 2\Delta\theta/k$. (iii) $\theta = 3\Delta\theta/k$. (iv) $\theta = 4\Delta\theta/k$. (v) $\theta = 5\Delta\theta/k$. (vi) $\theta = 6\Delta\theta/k$. (vii) $\theta = 7\Delta\theta/k$. (viii) $\theta = 8\Delta\theta/k$.

6.6. Classification results

We define the accuracy or correct recognition rate (CRR) by Eq. (19). The expression for CRR can be written as

$$CRR = \frac{N_c}{N_a} \times 100 \text{ (in percentage)} \tag{19}$$

where N_c is the total number of correct recognition sequences while N_a is the number of total action sequences. We use SVM for classifying actions. Table 2 shows the action recognition results of FBGDB where we use all global motion shape descriptors for each view. We use nine subjects, nine actions, and four views variation for testing (vA represents arbitrary view), and seven subjects, nine actions, and four views are used for training. We extract 51 features from SEI or each VT. The training and testing samples are selected randomly for several times. We perform the experiment by changing the training and testing samples randomly in each time. The training sets and the testing sets are different. As can be seen, there is a clear separation among different kinds of actions. The overall CRRs of $v1$, $v2$, $v3$, and vA are 79.34, 84.12, 90.47 and 82.53 respectively, of FBGDB. We use CVTs 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 eight subjects, six actions, and four scenarios and each scenario contains two or three action sequences and seven subjects, four actions, and four scenarios are used for training. Table 3 shows the recognition of each action in various scenarios for all the global shape descriptors. The CRRs of $s1$, $s2$, $s3$, $s4$, sA are 90.33, 84.17, 88.50, 89.3, 88.50 respectively for KTHDB, where sA is arbitrary scenarios.

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 variability models, and we make a comparison of performance among AT, VT, and CVT. As an example, the performance (in CRR) of AT, VT, and CVTs are 80%, 84.33%, and 90.33% respectively in case of FBGDB. Moreover, we show the system performance of KTHDB ($s1$ scenario only) without generating VTs. A comparison of performance is shown among ATs, VTs, and CVTs for different variations. Table 4 compares the recognition performance of each category. Therefore, the performance of proposed CVTs is significantly better than ATs.

Moreover, we present the performance of each feature set on test samples which is shown in Table 5. Different feature sets contribute in different proportions. The partial global motion distributions show the good performance. The combined features considerably improve the performance and characterize the shape geometry in multiple view-points.

6.7. Comparison

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 [4,8,9,13,15] 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

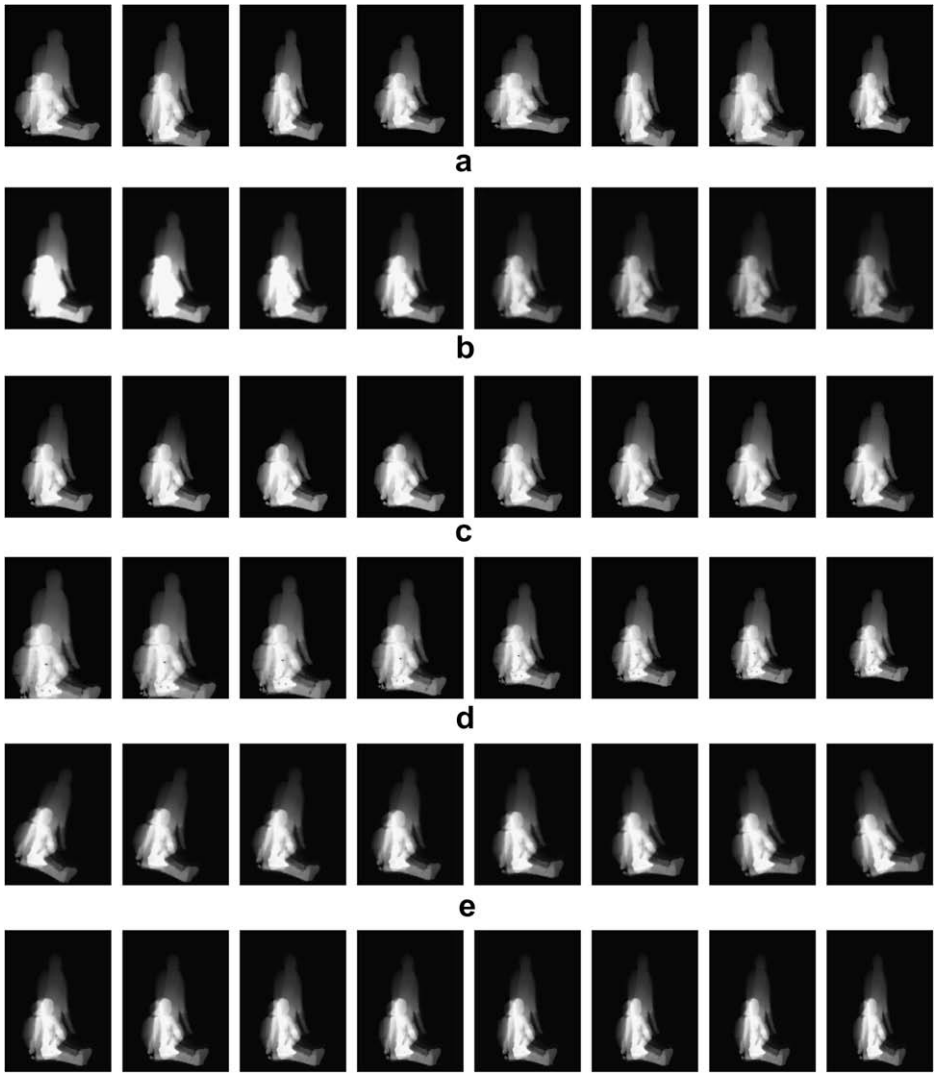

Fig. 13. Variability models of a typical action. (a) Anthropometric variation image for adaptation. (b) Speed variability images. (c) Starting (ϕ_s) and ending (ϕ_e) phase variation models. (d) Zoom variability models. (e) Variation models of person's slanting position. (f) Rotation variability images.

Table 2
CRRs of each action and each view of FBGDB.

	SC	WP	RH	BW	SF	GF	LF	RF	WF
*v*1	1.00	0.57	0.86	1.00	0.71	0.57	1.00	0.57	0.86
*v*2	0.86	0.43	1.00	0.86	0.86	0.86	1.00	0.86	0.86
*v*3	1.00	0.86	1.00	1.00	0.86	1.00	1.00	0.57	0.71
*v*A	0.95	0.57	0.95	1.00	0.76	0.76	0.90	0.76	0.76

Table 3
Recognition of actions of KTHDB.

Scenario	BP	HC	HW	JP	RP	WP
*s*1	1.00	0.97	0.97	0.91	0.74	0.83
*s*2	1.00	0.89	0.94	0.83	0.61	0.78
*s*3	1.00	0.97	0.98	0.74	0.81	0.81
*s*4	0.95	0.94	0.93	0.81	0.67	0.80

space time features. The overall comparison of different methods is listed in Table 6.

Compared to the mentioned researches, our approach yields best recognition results.

6.8. Limitations of the method

The possible limitations of the current method include

- We assume that the silhouette images are correctly extracted. But, we have limitations in extracting silhouette image properly.
- Silhouette energy image depends on the silhouette image. If the silhouette image is not correctly captured, them SEI becomes incorrect.
- In front view, all actions are not correctly recognized.
- Direction of actions is not possible to determine.

7. Conclusions and further research

This paper proposed a novel method for human action recognition using the SEI with variability action models. The variability provided a more natural and robust environment for human action recognition, using an advanced human–machine interface due to consideration on the following invariance factors: shape of actors, the starting and ending state of action, speed of an action, camera observations, and different scenarios. From the combined information of SEI and VTs, the global motion properties were extracted.

Table 4
Performance comparison of AT, each VT and CVT.

Actions	AT	$s(x,y)\|_a$	$s(x,y)\|_s$	$s(x,y)\|_\tau$	$s(x,y)\|_c$	CVT
BP	1.00	1.00	0.95	1.00	0.98	1.00
HC	0.88	0.96	0.90	0.95	0.96	0.97
HW	0.89	0.96	0.89	0.95	0.95	0.97
JP	0.78	0.88	0.82	0.84	0.82	0.91
RP	0.59	0.70	0.66	0.70	0.72	0.74
WP	0.66	0.75	0.70	0.76	0.75	0.83

Table 5
Performance of each kind of feature set ($s1$ scenario).

Input features	No.	CRR
Geom. and Zernike moments (s_g, s_z)	14	0.793
GM dist. and orien. (v_x, v_y, v_h, s_o)	5	0.743
Quadrant motion (s_{k_q})	4	0.815
Elliptical motion distribution (s_{k_e})	8	0.802
Partial block-motion (s_{k_b})	20	0.817
Overall ($s_g, s_z, v_x, v_y, v_h, s_{k_p}, s_o$)	51	0.905

Table 6
Comparison results of the action recognition.

Method	CRR	Scenario
Dollár et al. [4]	81.17	All
Jiang et al. [8]	84.43	All
Ke et al. [9]	62.96	All
Niebles et al. [13]	81.50	All
Schüldt et al. [15]	71.72	All
Our method	88.50	All
Schüldt et al. [15]	62.33	$s2$
Our method	84.17	$s2$

We recognized different daily human actions successfully in the indoor environment as well as outdoor environment. Moreover, by adapting the variability, incomplete actions were recognized successfully. The action recognition rate was not extremely higher but it was shown that we recognized actions from any view rather than a set view. At first, we tested our approach using FBGDB and then we performed experiments on KTHDB. We used the MCSVM classifier for performing classifications.

With adding the variability of human action, the action recognition becomes sparse and flexible for including multiple observations of human actions and it can be adapted to practical applications of human movement, human action recognition, and so on. This is in contrast to other individual shape based or motion-based variation approaches, as it includes both approaches. Moreover, by detection of the period or duration of an action from the image sequence, we need not consider all frames in the image sequence.

Despite of robustness, we face the problem of recognition in actions that occur at any view, as stated in the experimental section. This happens when two or more actions has similar body silhouettes maps and the global motion variation is low over the long image frames. Moreover, in the FBGDB, elderly peoples' movements are used for action modeling and recognition; and naturally show similar movement for similar actions. Due to 2D representation of human actions, the current limitation of our experiment is to compare the direction recognition, for example, sitting on a chair, or standing from a chair, or sitting on a chair and then standing, show the same recognition result, since 'sitting on a chair is the reverse way of 'standing from the chair' and the overall distribution of first half image sequence is similar to later half sequence. Our future work will include the precise detection and recognition of action with direction indication, improvement of all modules including adaptability. We will implement fully automated action recognition system and recognition of complicated parallel actions.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2009-0060113). This research was also supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea. This work was also supported in part by CASR grants KUET, Khulna, Bangladesh.

References

[1] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transaction on Circuits and Systems for Video Technology 18 (11) (2008) 1473–1488.
[2] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Transaction on PAMI 23 (3) (2001) 257–267.
[3] S. Carlsson, J. Sullivan, Action recognition by shape matching to key frames, in: IEEE Workshop on Models vs. Exemplars in CV, 2002, pp. 263–270.
[4] P. Dollár, G.C.V. Rabaud, S. Belongie, Behavior recognition via sparse spatio-temporal filters, in: IEEE Workshop VS-PETS, 2005, pp. 65–72.
[5] FBGDB. <http://gesturedb.korea.ac.kr/>.
[6] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transaction on NN 13 (2) (2002) 415–425.
[7] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Transaction on Information Theory 8 (1962) 179–187.
[8] H. Jiang, M.S. Drew, Z.N. Li, Successive convex matching for action detection, CVPR 2 (2006) 1646–1653.
[9] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, ICCV (2005) 166–173.
[10] A. Khotanzad, Y.H. Hong, Invariant image recognition by zernike moments, IEEE Transaction on PAMI 12 (5) (1990) 489–497.
[11] KTHDB. <http://www.nada.kth.se/cvap/actions/>.
[12] O. Masoud, N. Papanikolopoulos, Recognizing human activities, AVSBS (2003) 157–162.
[13] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, BMVC 3 (2006) 1249–1258.
[14] V. Parameswaran, R. Chellappa, View invariants for human action recognition, CVPR 2 (2003) 613–619.
[15] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, ICPR 3 (2004) 32–36.
[16] Y. Sheikh, M. Shah, M. Shah, Exploring the space of a human action, ICCV (2005) 144–149.
[17] C.J. Lin, R.C. Weng, Simple Probabilistic Predictions for Support Vector Regression, Technical Report, National Taiwan University, 2004.
[18] J. Westons, C. Wtkins, Support vector machines for multiclass pattern recognition, in: Proceedings of the 7th European Symposium on Artificial Neural Networks, 1999, pp. 219–224.
[19] M. Ahmad, S.-W. Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (2008) 2237–2252.
[20] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Transaction on PAMI 28 (2) (2006) 316–322.
[21] T.H.W. Lam, R.S.T. Lee, A new representation for human gait recognition: motion silhouette image, LNCS 3832 (2006) 612–618.