Human Action Recognition from Variable Silhouette Energy Images

by

Irine parvin

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Engineering in Electrical and Electronic Engineering

Khulna University of Engineering & Technology
Khulna 9203, Bangladesh.

June 2011

Declaration

This is to certify that the thesis work entitled "Human Action Recognition from Variable Silhouette Energy Images" has been carried out by Irine Parvin in the Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.

Signature of Supervisor

Irine Parvin

Signature of Candidate

Approval

This is to certify that the thesis work submitted by Irine Parvin entitled "Human Action Recognition from Variable Silhouette Energy Images" has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of Master of Engineering in the Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh in June 2011.

BOARD OF EXAMINERS

1. 14.06.2011 ______________________ Chairman (Supervisor)

Prof. Dr. Mohiuddin Ahmad
Dept. Electrical & Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203

2. Rafiq 14.06.11 ______________________ Member

Prof. Dr. Md. Abdur Rafiq
Head of the Department
Dept. Electrical & Electronic Engineering
Khulna University of Engineering & Technology

3. ______________________ Member

Mr. A. N. M. Enamul Kabir
Associate Professor, Dept. Electrical & Electronic Engineering
Khulna University of Engineering & Technology

4. 14/06/11 ______________________ Member

Mr. Md. Salah uddin Yusuf
Assistant Professor, Dept. Electrical & Electronic Engineering
Khulna University of Engineering & Technology

5. 14/06/11 ______________________ Member (External)

Prof. Dr. Md. Al-Amin Bhuiyan
Dept. of Computer Science & Engineering
Jahangirnagar University, Savar, Dhaka

ACKNOWLEDGEMENT

I thank to almighty ALLAH for successful completion of my thesis.

I would like to express my heartfelt gratitude to my supervisor Prof. Dr. Mohiuddin Ahmad to accept me as graduate student and to provide a modern and efficient research environment. I am grateful for his ceaseless guidance, efficacious suggestions and constant motivation throughout my entire course studies and thesis work. His gracious attention helps me in building confidence to do my research. Prof. Dr. Mohiuddin Ahmad is an intellectual in recognizing human actions. He is the source of vast ideas and profound knowledge, and feedback all the time for authors. Due to these helpful qualities the author gratefully acknowledges her profound indebtedness to him.

I would like to thank Prof. Dr. Md. Rafiqul Islam, Dean, Faculty of Electrical & Electronic Engineering, KUET for his advice and inspiration to work under Prof. Dr. Mohiuddin Ahmad.

Furthermore, I would like to thank all the faculty members of Electrical & Electronic Engineering Department, KUET for helping in various ways in formulating this dissertation.

Finally, I would like to thank my parents for being the ultimate source of strength and encouragement throughout my academic career and personal life and particularly to my husband who has always supported my education with his love and encouragement.

Abstract

Recognizing human actions is an important issue in the computer vision community. Human action recognition becomes more challenging when variability areas such as, anthropometric variation, phase variation, speed variation, camera view variation, individual variations in appearance and clothes of people, changes in light and view point and so on. In this thesis, we propose a spatio-temporal silhouette representation, 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. We also proposed variable silhouette energy image for different variable situations. To address the variability in the recognition of human actions several parameters such as, anthropometry of person, phase (starting and ending state of action) speeds of the actions, camera observation (distance from camera, tilting motion and rotation of human body) and view variations are proposed. The SEI and SHI are constructed using the silhouette image sequence of an action. The span or difference of the end time start time is used to make SHI. We extract the features based on geometrical shape moments. Using the features, we generate a unified description of model by learning the multi-class SVM for each action. Finally we recognize action using action model for any arbitrary image sequence. We tested our approach successfully in the indoor and outdoor environment. Our experimental results show that the proposed method is robust, flexible and efficient.

Contents

LIST OF TABLES

LIST OF FIGURES

Abbreviations

AT	Action templates
CLG	Combined local global
COM	Center of mass
CRR	Correct recognition rate
GMD	Global motion description
SEI	Silhouette energy image
SHI	Silhouette history image
SVM	Support vector machine
VT	Variability template

Nomenclature

$S(x,y)$	Silhouette energy image
$\sigma(x,y)$	Standard deviation image
$Cv(x,y)$	Motion variation
t_s	Starting state of an action
t_e	Ending state of an action
α_t	Temporal coefficient
d_t	Temporal constant
$H_t(x,y,t)$	Silhouette history image
τ	Duration of temporal extension to previous silhouette image
s	Person's velocity
N	No. of frames are needed to perform the action
ϕ_s	Starting phase
ϕ_e	Ending phase
R	Diameter of a cylinder
ρ	Radius of the cylinder
M_{pq}	Cartesian moment of order (p+q)
m_{pq}	Two dimensional Cartesian moment
μ_{pq}	Two dimensional centralized moment
n_{pq}	Two dimensional scale-normalized centralized moments
Z_{nm}	Two-dimensional Zernike moments
M_c	Covariance matrix
λ_i	Eigenvalue of the covariance matrix
N_c	No. of correct recognition
N_a	No. of total recognition

CHAPTER 1

Introduction

Human action recognition has been running a fertile topic of study for several years and there exists a vast body of literature on the subject. It has potential application in video surveillance and monitoring, human-computer interaction, human robot interaction, model based compression, augmented reality etc. The existing methods of human action recognition can be categorized depending on the image state properties such as, motion based, shape based, gradient based etc. Researchers either used an explicit body shape model or did not use any body shape model for action recognition with and without human motion information. Our approach can be considered as a combination of shape and motion based representation without using any prior body shape model.

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 answer the following question. Which feature is robust for action recognition in critical conditions or varying environment? Usually, 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, and so on, and the motion of the human body is non-rigid in nature. These characteristics make human action recognition as a more challenging and sophisticated task. Considering the above circumstances, we consider some issues that affect the development of models of actions and classifications, which are as follows:

- 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 changes, which are shown in Fig. 1.1[18]

- An action can be viewed as a series of silhouette images of the human body (Fig. 1.1(b)). The silhouette information involves no translation, rotation, and scaling. Moreover, the silhouette sequence of an action is invariant to the speed.

- Action can be viewed by the motion or velocity of human body parts (Fig. 1.1(c)). Simple action involves the motion of a small number of body parts and complex action involves the motion of a whole body. The motion is non-rigid in nature.

- Human action depends on anthropometry, method of performing the action, phase variation (starting and ending time of the action), scale variation of an action, and so on.

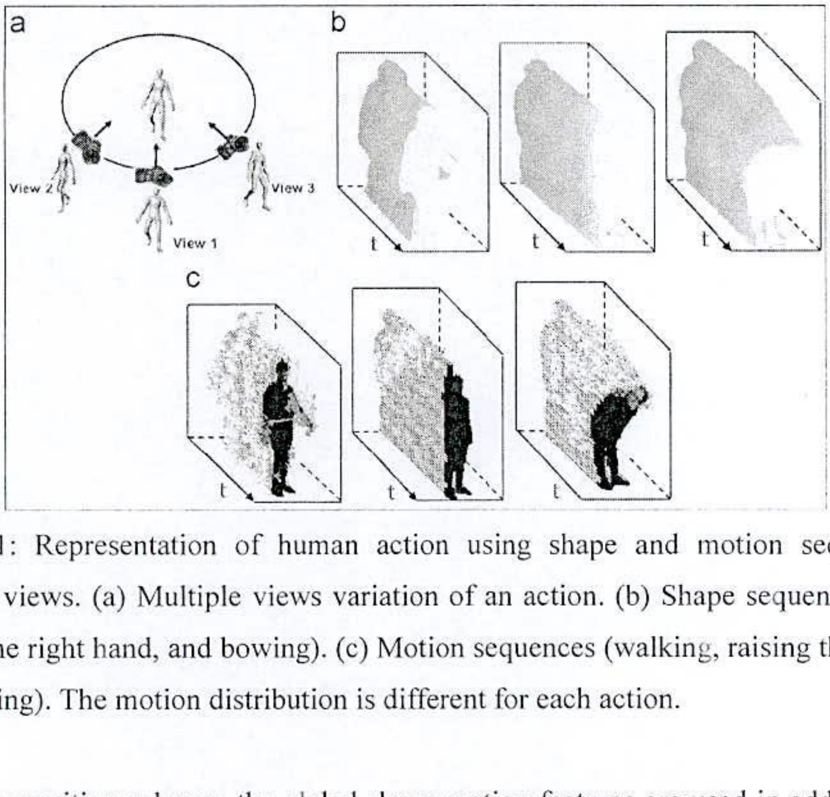


Figure1.1: Representation of human action using shape and motion sequences with multiple views. (a) Multiple views variation of an action. (b) Shape sequences (walking, raising the right hand, and bowing). (c) Motion sequences (walking, raising the right hand, and bowing). The motion distribution is different for each action.

In this recognition scheme, the global shape motion features are used in addition to some variables for recognizing the periodic as well as non-periodic actions.

1.1 Motivation

Our work is motivated by the ability of human to utilize periodic and no periodic motion to perform various actions. Many actions are periodic in nature. These periodic natures of human actions can be analyzed using the shape of human beings, such as shape 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 human describes the nature of human's motion and shows the action or activity performed by human. 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 relational characteristics of each action.

1.2 Objective

The main objective of this thesis is to recognize daily life human action such as walking, running, jogging, and sitting on a chair, hand raising, sitting on the floor, getting down on the floor, etc. by using global shape motion descriptions from the image sequences with selected variability. In order to address the variability in the recognition of human actions, several parameters such as anthropometry of person, phase (variation of starting and ending state of an action), camera observations (zooming of the person. tilting of the person and rotation of the person) and camera view variation are proposed. From the research we can recognize different daily life human actions in the indoor and outdoor environment.

1.3 Contribution

In this thesis, image similarity is used to recognize human actions. The major contributions are as follows:

1. Variable silhouette energy images are used for action representations. These representations are used for extracting global features for recognizing specific human actions.

2. Shape variation of image sequence in time is represented here in this thesis by using SHI (Silhouette history image). It shows not only the motion images representation but also the global motion orientation of an action at any instant of time.

3. Explicit variability action model is introduced for considering different forms of 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. In general these explicit variability action models are represented as variable SEI.

1.4 Application

In general, the human action recognition from multiple view-points has the applications in video surveillance and monitoring, human computer interactions, model based compressions and video retrieval in various situations etc. We recognize periodic and non-periodic human action in different scenarios such as scale variations, appearance and clothing variations, arbitrary views and incomplete actions. We have tested our proposed approach on the Korea university gesture database (KUGDB) [16], which represents key actions in daily life of elderly persons and KTH action database (KTHDB) [17] which represents the multiple scenarios action data. We recognize several daily actions of elderly people for human robot interaction (HRI) or similar applications.

1.5 Overview of the system

The proposed method is shown in fig. 1.2. The operation is divided into two phases, such as learning phase and classification or recognition phase. In this method, at first the action area (spatial and temporal boundary) is searched and then the duration of an action is estimated. Depending on the temporal boundary, SEI & SHI are constructed from silhouette image. Variability models are generated by using variable parameters. The variable parameters are anthropometry, speed, phase, camera observations. From all the models, we extract some salient features. Using the features a unified description of the model was generated by learning the SVM for each action. Finally action can be recognized using action model for any arbitrary image sequence.

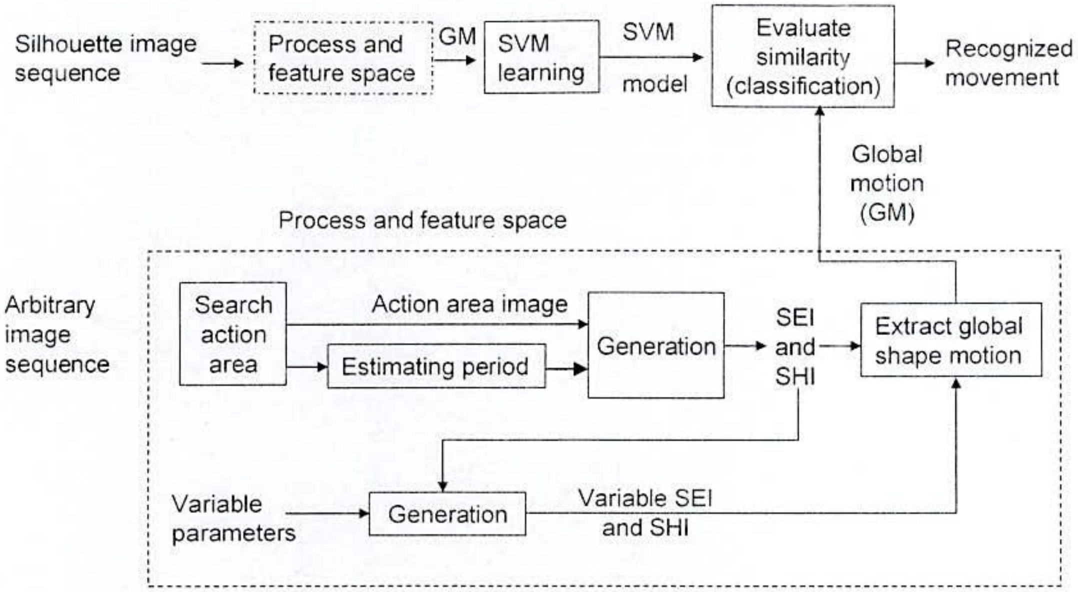


Figure 1.2: Illustration of the proposed approach of human movement recognition system.

In this method there are some assumptions, such as:

- In the case of silhouette image, it is assumed that silhouette images are correctly extracted. If silhouette images are not correctly extracted then the accurate result can not be got.
- The duration of each action is assumed a few seconds. As many actions have to be counted, the duration of actions is taken more than few seconds; otherwise it will be difficult to calculate. Since we use the number of frames for making SEI as SHI of gray level image, therefore approximately 255 image frames are used.
- The camera position is stationary, only the person moves within the camera view. Here we have considered camera observation variation (zoom, till and rotation).
- It is considered maximum 15% partial occlusion on the person. More occlusion decrease the accuracy.
- The brightness of an image sequence will be constant. Change of brightness means the change of SEI and SHI. Change of SEI and SHI means change of features.

CHAPTER 2

Related works

Human action recognition from human image sequences is an active area of research in computer vision. It is simply a classification problem involving time series feature data. Recognition consists of matching an unknown test sequence with a library of labeled sequences that represent the prototypical actions. Several surveys have attempted to classify various approaches for solving the problem. Detailed surveys can be found in [1-4], where different methodologies of human action recognition, human movement, etc. are discussed.

2.1 Shape-based Approach

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 [5-8,]. The authors [5] recognized complex motion features patterns by using local space time features and integrated these representations with SVM classification schemes for recognition. They used variation in scale, the frequency and velocity of pattern. For purpose of evaluation they introduced a new video database containing 2391 sequences of six human actions performed by 25 people in few scenarios. Here it is proposed the anthropometry variation, speed and view variation, and view observation. The KTHDB database is considered in this work which contains six types of human actions, performed several times by 25 subjects in four scenarios and 2391 sequences. Doll´ar et al. [7] proposed the approach to detect sparse space-time interest points based on separable linear filters for behavior recognition. Niebles et al. [8] used local space time features for unsupervised learning of human actions. They presented a "Bag of video words" model combined with space time interest point detector, for human action categorization and localization. They localized multiple actions in complex motion sequence containing multiple actions.

Yan and Rahul [6] used volumetric features which was an alternative to popular local descriptor approaches for event detection in video sequences. They generated the notion of 2D box features to 3D spatio temporal volumetric features. They varied the view pointy, scale and action speed. [9] H. Jiang presented a successive convex programming scheme to match video sequences using intra-frame inter frame constrained local features. By convexifying the optimization problem sequentially with an efficient linear programming scheme which can be globally optimized in each step gradually shrinking the trust region, their method was more robust than previous matching scheme. Masoud [10] approached a motion recognition which was based on low level motion features computing by using an IIR filter.

2.2 Motion-based Approach

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 [11] [12]. Bobick and Davis [11] 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. Carlsson and Sullivian [12] demonstrate the specific posture that can be recognized in long video sequence by matching the shape information extracted from individual frames to store prototypes representing key frames of the action. The matching algorithm is tolerant to substantial deformation between image and prototype qualitatively similar image shape produced by the body postures.

2.3 Shape and Motion based Approach

In addition of shape and motion, several variabilities that occurred frequently are also responsible for human action recognition. Sheikh and Shah [13] 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. Yilmaz el al. [15] Performed action recognition in the presence of camera motion.

M. Ahamd and S. Lee [19] have presented a method for human action recognition from multi-view image sequences that use the combined motion and shape flow information with variability consideration. A combined local–global (CLG) optic flow is used to extract motion flow feature and invariant moments with flow deviations which are used to extract the global shape flow feature from the image sequences. In our approach, human action is represented as a set of multidimensional CLG optic flow and shape flow feature vectors in the spatial–temporal action boundary. Actions are modeled by using a set of multidimensional HMMs for multiple views using the combined features, which enforce robust view-invariant operation. They recognize different human actions in daily life successfully in the indoor and outdoor environment using the maximum likelihood estimation approach. The results suggest robustness of the proposed method with respect to multiple view action recognition, scale and phase variations, and invariant analysis of silhouettes.

2.4 View based Approach

Most of the human action recognition techniques depend on the viewing direction. However, the trajectory of an action from different viewing angles is different. The work of testing an action using multi-view motion learning is not well resolved. Seitz and Dyer in Ref. [23] described an approach to detect cyclic motions that is affine invariant. Rao and Shah [24] again used view invariant actions by affine invariance assuming that 2D positions of the hand are already known. This approach utilized spatiotemporal curvature maxima as instants of interest to map an unknown viewpoint to a "normal" viewpoint. The action was considered as being completely represented by the motion of the hand alone. In Ref. [25], authors presented human action in video using 3D model-based invariants and represent each action using a unique curve.

2.5 Our proposed Approach

In [26], we proposed action models by using silhouette energy image (SEI) and silhouette history image (SHI). We also generate variable energy images using SEI and several control parameters, such as anthropometry variation, speed variation and camera observation variation.

In [27], we have presented a method of human action recognition that uses spatial-temporal local and global motion descriptions from the image sequence with selected variability. In this method, the motion of the body parts, i.e., local motions are extracted from the image sequence by using optical flow and global motions are extracted by using principal components of each action. Flow and the component features are added with mentioned variabilities for recognizing periodic as well as non-periodic action. Support vector machine is used for learning and recognizing actions.

CHAPTER 3

Human Action Model Representations

3.1 Definition of Action

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. It is called Periodic action such as walking, running, jogging etc. Examples of periodic actions are shown in Fig 3.1. Besides, many actions show single occurrence or non-periodicity with time frame of specific length (i.e. duration). That is called non-periodic action such as raising hand, bowing, sitting on the floor, bending, jumping etc. Examples of non-periodic actions are shown in Fig 3.2.

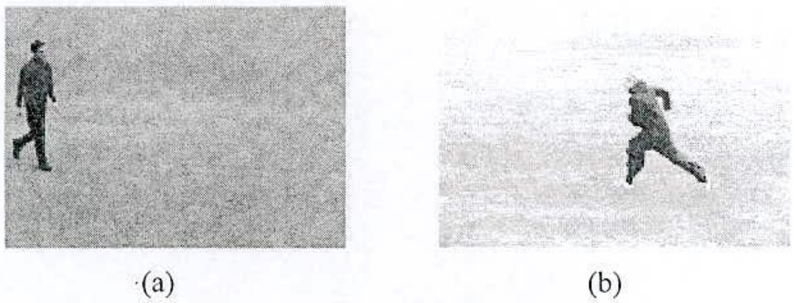

(a) (b)

Figure 3.1: Examples of periodic action (a) Walking at a place (b) Running at a place

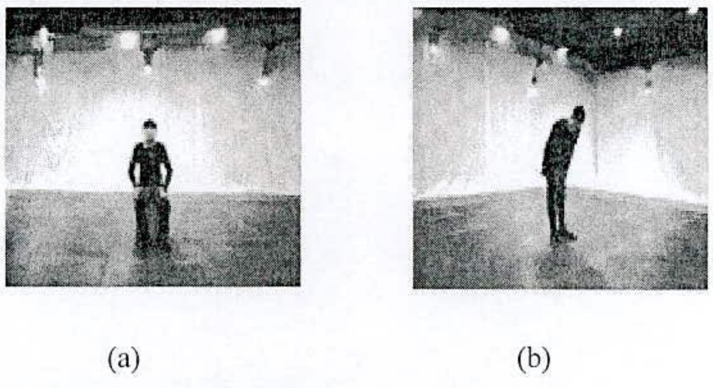

(a) (b)

Figure 3.2: Examples of non-periodic action (a) Sitting on a chair (b) Bowing

3.2 Actions in World Coordinate System

It is considered that a complete human action representation might be the set of all three dimensional points on a performing actor. Therefore, human actions are considered as four dimensional points in real world-space, which can be represented as follows:

$$A_{4D} = \begin{pmatrix} X_j^{T_1} & X_j^{T_2} \dots X_j^{T_P} \\ Y_j^{T_1} & Y_j^{T_2} \dots Y_j^{T_P} \\ Z_j^{T_1} & Z_j^{T_2} \dots Z_j^{T_P} \end{pmatrix} \qquad (3.1)$$

where X, Y, and Z represent the state-space representation of a point in the 4D plane of a person performing an action. Here, j represents the point set, or anatomical landmarks points, or voxel and $_{Ti}$ is the i_{th} frame in the world coordinate. When the human action is projected into the spatial–temporal image plane, then equation (3.1) can be represented by

$$A_{3D} = \begin{pmatrix} x_j^{t_1} & x_j^{t_2} \dots & x_j^{t_P} \\ y_j^{t_1} & y_j^{t_2} \dots & y_j^{t_P} \end{pmatrix} \qquad (3.2)$$

where x and y represent the 2D points in the spatiotemporal image space or 3D space and j represents the point set of pixel in the action region and t_i is the i_{th} frame in the image coordinate in performing an action. The relation between t and T could be linear, such that $t = \alpha_t T + d_t$, where α_t and d_t represent the temporal coefficient and temporal constant, respectively

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.

3.3 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. 3.3 (a). Therefore, SEI (SEI $=S(x, y)$) is defined by (3.3).

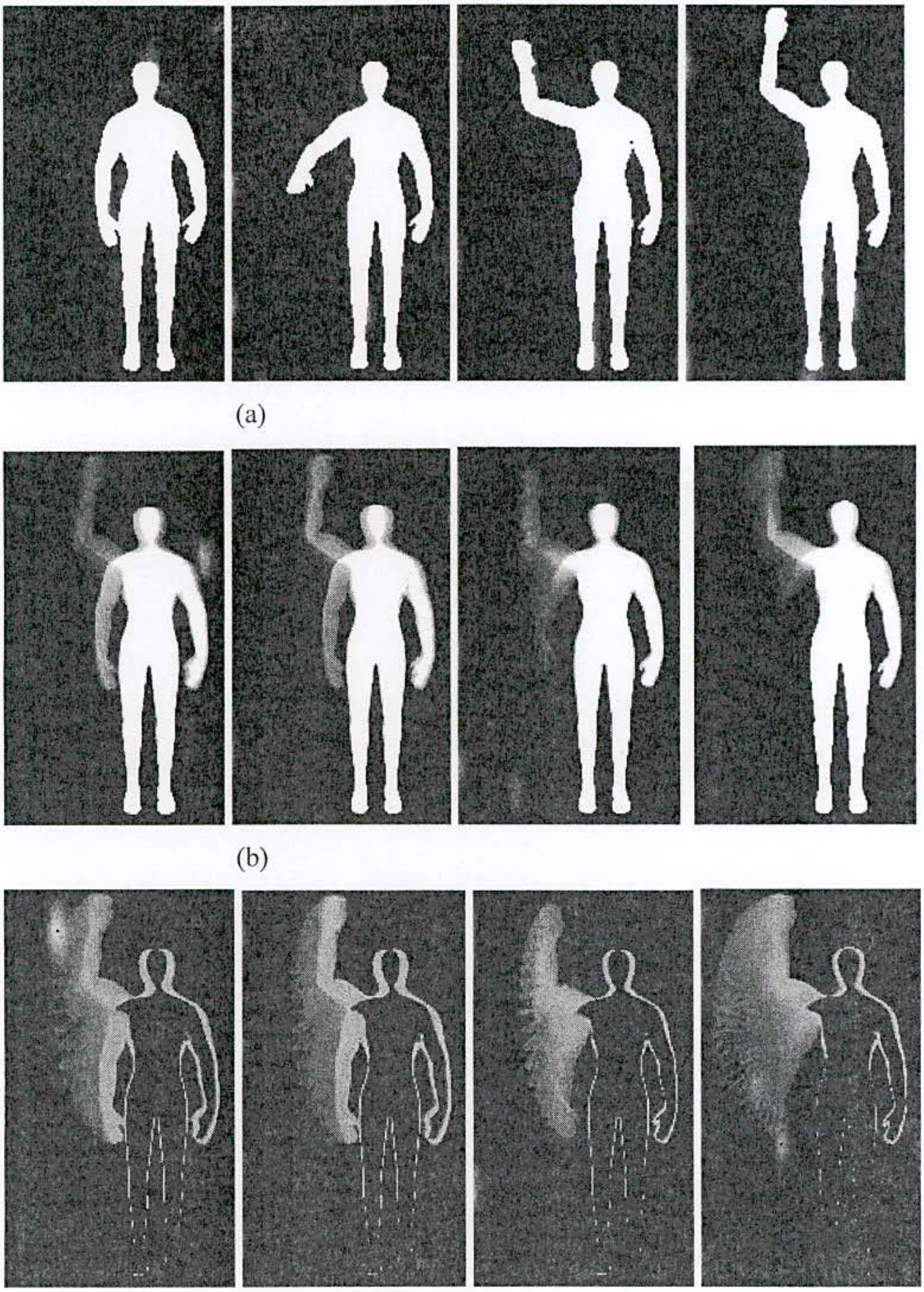
Moreover, the standard deviation image as well as motion variation expressions are given by (3.4) and (3.5).

$$S(x,y) = \frac{1}{t_e - t_s} \int_{t_s}^{t_e} x_t \, dt \tag{3.3}$$

$$\sigma(x,y) = \sqrt{\frac{\int_{t=t_S}^{t_e} x_t^2 \, dt}{t_e - t_s} - \left(\frac{1}{t_e - t_s} \int_{t=t_s}^{t_e} x_t \, dt \right)^2} \tag{3.4}$$

$$C_v(x,y) = \frac{\sigma(x,y)}{S(x,y)} \tag{3.5}$$

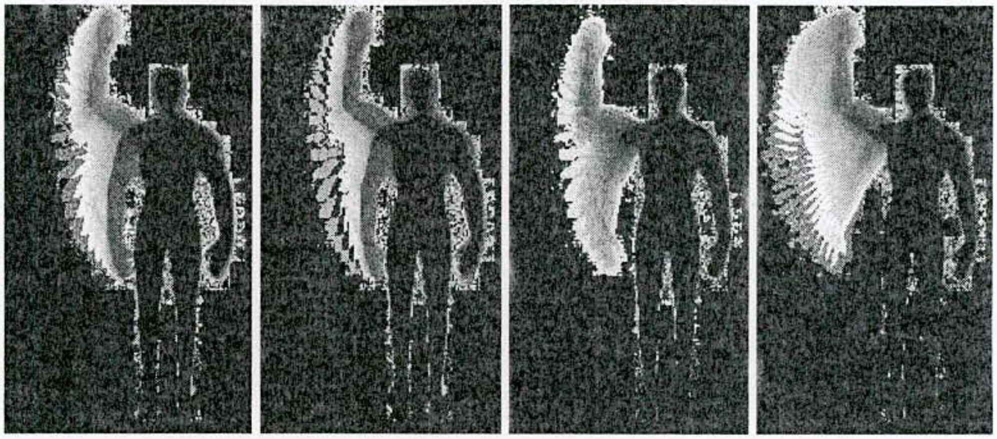
Here, t_s and t_e are the starting and ending states of an action. Alternately, we can represent total number of frame by F and hence the period is F. Therefore, the period or duration becomes $F = t_e - t_s$. Since the average 2D image stores the global motion distribution and orientation of the silhouette images, this can be designated as a SEI. The number of frames in the action depends on the person, time, and type of action. Since, the average of the time sequence silhouette images is used, the normalized variation affects are very low. Fig. 3.3(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. 3.3 (b), Fig. 3.3(c) and Fig. 3.3(d) for different spans of time, respectively. This representation shows the shape as well as motion changes of an action. The SEI represents an action model (AM), 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.



(a)

(b)

(c)



(d)

Figure 3.3: 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.

3.4 Justification of SEI

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 an 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. Now, the silhouette image is given by

$$x_t(x,y) = o_t(x,y) + n_t(x,y) \tag{3.6}$$

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

$$n_t(x,y) = \begin{cases} n_{lt}(x,y): & \begin{cases} P\{n_t(x,y) = -1\} = p \\ P\{n_t(x,y) = 0\} = 1-p \end{cases} if \quad o_t(x,y) = 1 \\ n_{2t}(x,y): & \begin{cases} P\{n_t(x,y) = 1\} = p \\ P\{n_t(x,y) = 0\} = 1-p \end{cases} if \quad o_t(x,y) = 0 \end{cases} \tag{3.7}$$

We have $$E\{n_t(x,y)\} = \begin{cases} -p: & if \quad o_t(x,y) = 1 \\ p: & if \quad o_t(x,y) = 0 \end{cases} \tag{3.8}$$

and $$\sigma^2_{n_t(x,y)} = \sigma^2_{n_{lt}(x,y)} = \sigma^2_{n_{2t}(x,y)} = p(1-p) \tag{3.9}$$

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

$$S(x,y) = \frac{1}{F}\sum_{t=1}^{F} x_t(x,y) = \frac{1}{F}\sum_{t=1}^{F} o_t(x,y) + \frac{1}{F}\sum_{t=1}^{F} n_t(x,y) = \frac{G}{F} + \overline{n}(x,y) \tag{3.10}$$

Now, the noise in silhouette image is

$$\overline{n}_t(x,y) = \frac{1}{F}\left[\sum_{t=1}^{G} n_{1t}(x,y) + \sum_{t=G+1}^{F} n_{2t}(x,y)\right] \tag{3.11}$$

The first moment is given by

$$E\{\overline{n}_t(x,y)\} = \frac{1}{F}\left[\sum_{t=1}^{G} E\{n_{lt}(x,y)\} + \sum_{t=G+1}^{F} E\{n_{2t}(x,y)\}\right] = \frac{1}{F}\left[G(-p) + (F-G)p\right] = \frac{(F-2G)p}{F} \tag{3.12}$$

And the variance,

$$\begin{aligned} \sigma^2_{\overline{n}_t(x,y)} &= E\left\{\left[\overline{n}(x,y) - E\{\overline{n}(x,y)\}\right]^2\right\} \\ &= \frac{1}{F^2} E\left\{\left[\sum_{t=1}^{G} n_{lt}(x,y) - E\{n_{lt}(x,y)\}\right] + \left[\sum_{t=G+1}^{F}\left[n_{2t}(x,y) - E\{n_{2t}(x,y)\}\right]\right]^2\right\} \\ &= \frac{1}{F^2}\left[G\sigma^2_{n_{lt}(x,y)} + (F-G)\sigma^2_{n_{2t}(x,y)}\right] = \frac{1}{F^2}\sigma^2_{n_t(x,y)} \end{aligned} \tag{3.13}$$

Therefore, the mean of the noise in SEI varies between $-p$ and p depending on G while its variability (second moment of noise) decreases.

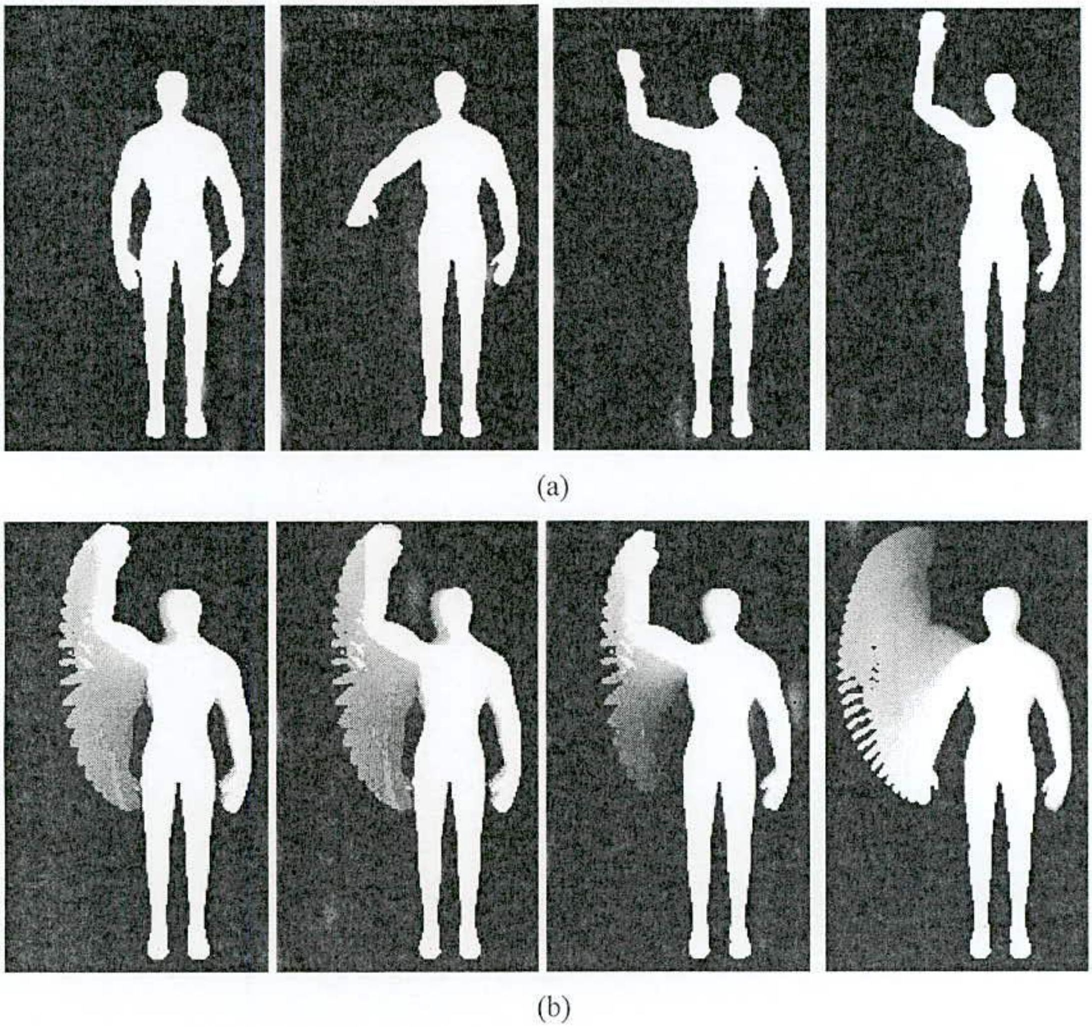
If $G = F$ at (x, y)(all $o_t(x, y) = 1$), then $E\{\overline{n}_t(x, y)\} = -p$. If $G = 0$ at (x, y), then (all $o_t(x, y) = 0$). At the position (x, y) the mean of the noise in SEI is the same as that in individual silhouette image, but the noise variance reduces so that the probability of outliers is reduced. If $G = 0 \sim F$ at (x, y), $E\{\overline{n}_t(x, y)\} = p \sim -p$. Therefore both the mean and the variance of noise in SEI are reduced compared to individual silhouette image at these locations. All 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.

3.5 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 only shows the motion image representation as [11] but also 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. 3.4. 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 [11], 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 [11]. We use silhouette images instead of motion images for making a SHI given by (3.14).

$$H_t(x, y, t) = \begin{Bmatrix} \pi, \ \mathit{if} \ x_t(x, y) > 1 \\ \max(0, H_t(x, y, t-1)), \mathit{otherwise} \end{Bmatrix} \tag{3.14}$$

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 timè t. SEI and SHI are regarded as the action models (AMs) in our approach.



(a)

(b)

Figure 3.4. Human action representation using silhouette history image (SHI) (a) Some key frames of an action (b) SHI at the same time span of (50 frames, 60 frames, 20 frames, and 50 frames).

CHAPTER 4

Variable Silhouette Energy Images

4.1 What are variable silhouette energy image or variable action models?

The variable silhouette energy image or variable action models 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. Generation of SEI is shown in Fig. 4.1. We can say,

$$\text{Variable silhouette energy images} = f(\text{SEI, control parameters}) \tag{4.1}$$

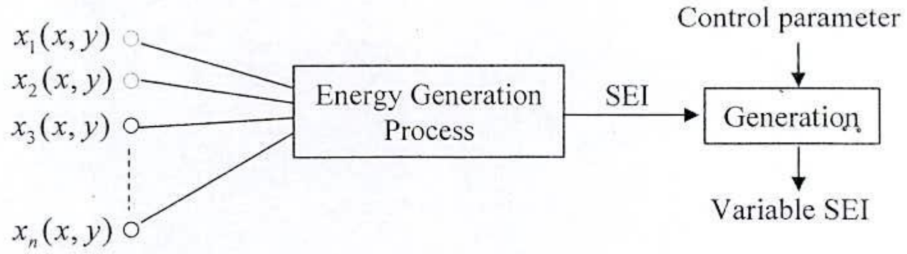


Figure 4.1 Generation of variable silhouette energy images

4.2 Types of Variability for human action recognition

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

- Anthropometry variation
- Speed variation
- Phase variation
- View observation and view variation

We define the variabilities of human action by following expression:

4.2.1 Anthropometry variation

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. 4.2 shows the example of anthropometry variation. Due to different girth and height variations, human action models should adapt anthropometry. 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. These variations are modeled using anthropometric variation. Due to the variation of human anthropometry as shown in Fig. 4.2,

$$S(x,y)\Big|_{adapt} = \{S(x,y)\Big|_{anthro,} S(x,y)\Big|_{speed,} S(x,y)\Big|_{phase,} S(x,y)\Big|_{camera}\} \tag{4.2}$$

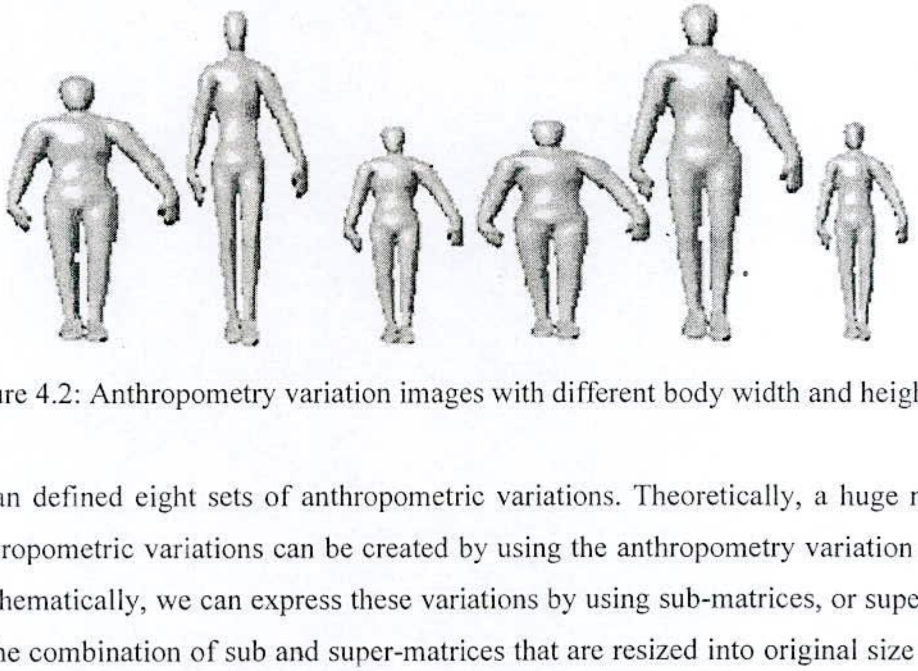


Figure 4.2: Anthropometry variation images with different body width and height.

It can defined eight sets of anthropometric variations. Theoretically, a huge numbers of anthropometric variations can be created by using the anthropometry variation parameter. 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.

$$S(x,y)\,|_{antro} = \begin{cases} S(x-a,y), & \text{Higher width} \\ S(x,y-b), & \text{Higher height} \\ S(x+a,y), & \text{Lower width} \\ S(x,y+b), & \text{Lower height} \\ S(x-a,y+b), & \text{Higher width Lower height} \\ S(x+a,y-b), & \text{Lower width Higher height} \\ S(x-a,y-b), & \text{Higher width Higher height} \\ S(x+a,y+b), & \underbrace{\text{Lower width Lower height}}_{\text{output}} \end{cases} \tag{4.3}$$

where, *a* and *b* are the parameters for maximum allowable anthropometric variations. The modeling of the parameters is not same for all types of action due to the motion of different body parts. We propose to construct the anthropometric adaptable models *S(x-a, y-b)* from a given original action model and anthropometric parameters. The other adaptable models follow the similar procedure of Fig. 4.4. Using the variation of anthropometry shown in equation (4.3), and by following Fig. 4.4, we can simulate the adaptable anthropometric models as shown in Fig. 4.3. In general, we can assume similar parameters for all kinds of action.

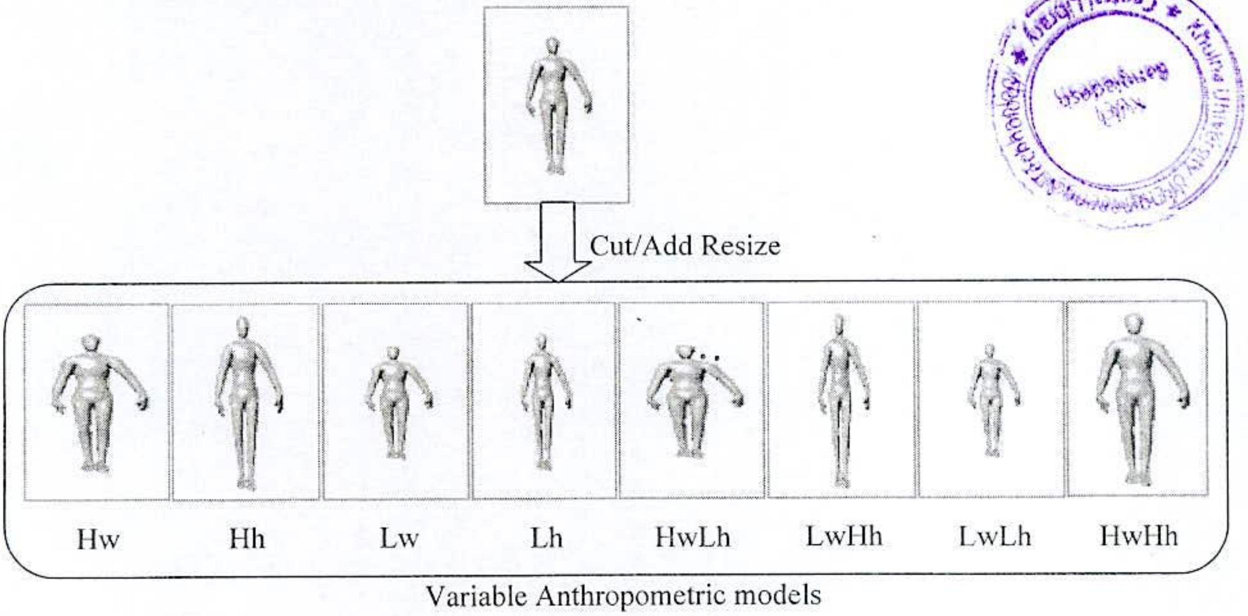


Figure 4.3: Variable Anthropometric models

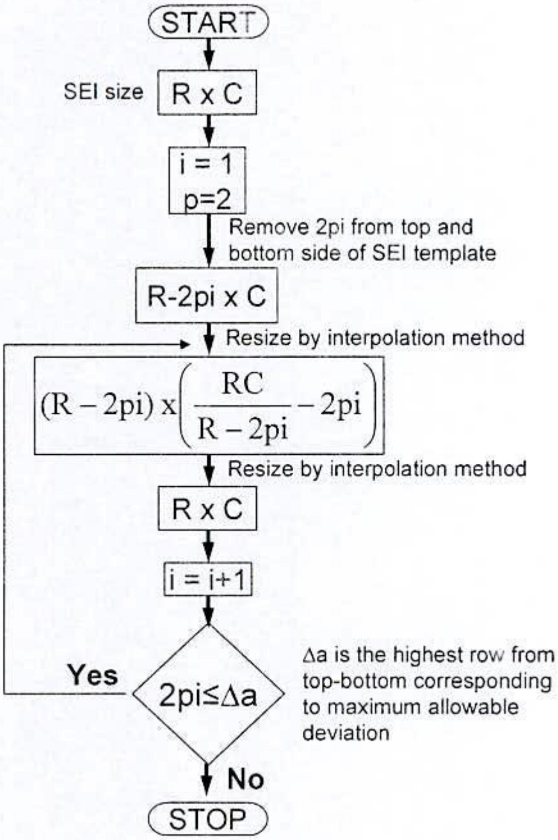

START
SEI size
R x C
i = 1
p=2
Remove 2pi from top and bottom side of SEI template
R-2pi x C
Resize by interpolation method
$(\mathrm{R}-2\mathrm{pi})\,\mathrm{x}\left(\frac{\mathrm{RC}}{\mathrm{R}-2\mathrm{pi}}-2\mathrm{pi}\right)$
Resize by interpolation method
R x C
i = i+1
Yes
2pi≤Δa
Δa is the highest row from top-bottom corresponding to maximum allowable deviation
No
STOP

Figure 4.4: Flow chart for generating anthropometric variable model.

4.2.2 Speed variation

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, 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.

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 $N >> n$. 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 x 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. (4.4). This does not rigorously follow the speed variation, but it approximates the variation of speed of action.

$$s(x,y)\big|_s = \begin{cases} s(x,y)\left(\dfrac{N}{N+n}\right)\dfrac{1}{2\pi\sigma^2}e^{\frac{x^2+y^2}{2\pi\sigma^2}}, \text{Actionperiod>}N \\ s(x,y)\left(\dfrac{N}{N-n}\right)\dfrac{1}{2\pi\sigma^2}e^{\frac{x^2+y^2}{2\pi\sigma^2}}, \text{Actionperiod>}N \end{cases} \tag{4.4}$$

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

4.2.3 Phase variation

The variable 'phase variation' refers to an action occurred at different starting and ending state. The starting and ending phases of an action depend 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. (4.5).

$$s(x,y)\big|_{\tau}=\begin{cases}\frac{1}{p-\phi_s}\sum_{t_{s+\phi_s}}^{t_e} x_t\,,\phi_s\,\text{varies}\\ \frac{1}{p-\phi_e}\sum_{t_s}^{t_{e-\phi_e}} \phi_e\,,\phi_e\,\text{varies}\end{cases} \tag{4.5}$$

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 change because of few frames blank (which we can consider incomplete actions). An illustrating situation of phase variation is shown in Fig. 4.5.

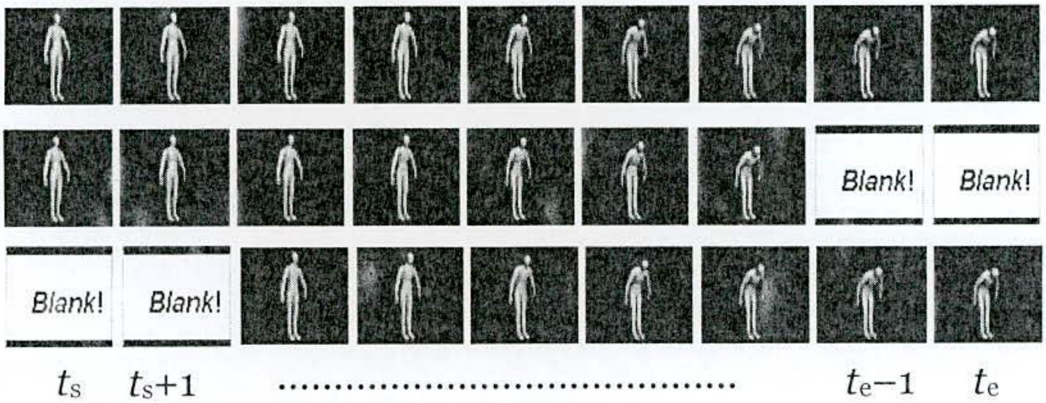


Figure 4.5: Phase variation. Top row: complete action. Middle and bottom row: incomplete action. ''Blank!" refers to some frames missing at start or end. Here, t_s and t_e are the starting and ending state of the action.

4.2.4 Camera parameter observations

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 parameter observations 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. 4.6 illustrate the camera observations 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 affinitely 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. (4.6).

$$S(x,y)|_c = S\left(\begin{bmatrix} d+S_x & S_y - r \\ r+S_y & d-S_x \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}\right) \tag{4.6}$$

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. 4.6.

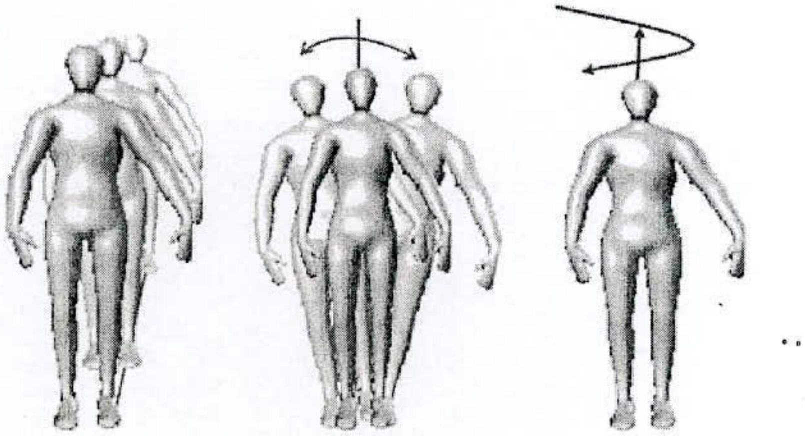


Figure 4.6: Observation and view variation of human actions. (a) Camera observations. Left: person's distance from camera (scaling of a person). Middle: slanting position of human body. Right: human body rotation around upward axis.

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

$$f_{c.rot}(x,y,t) = f(x + 2\rho(1-\cos\theta), y, t) \tag{4.7}$$

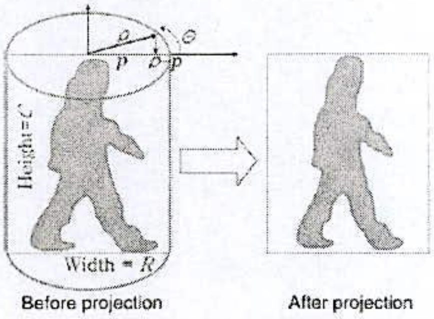


Figure 4.7: Illustration of 2D projection geometry.

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

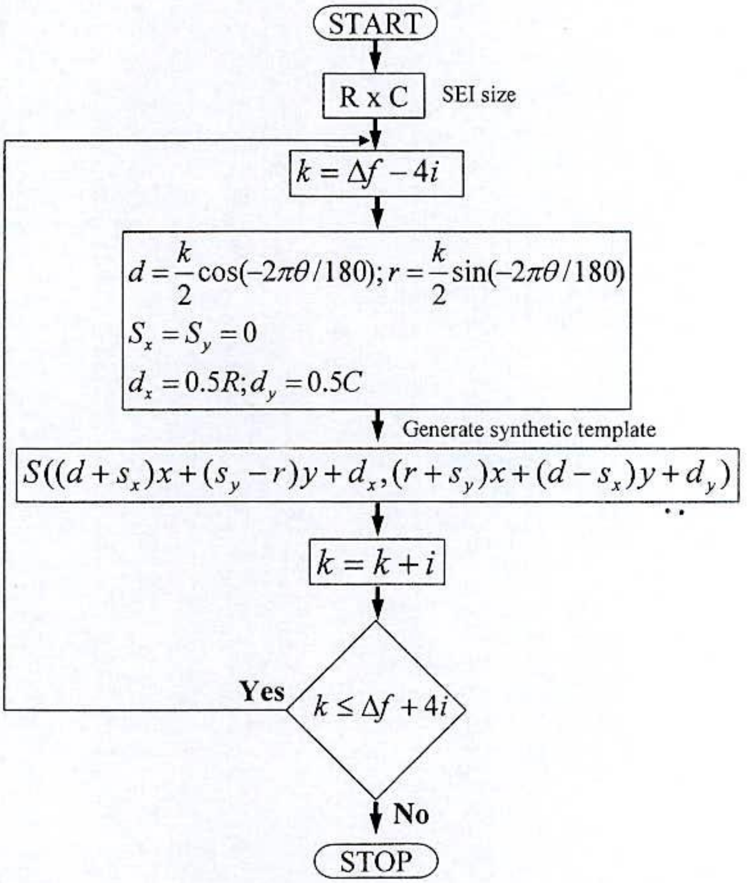

START
R x C
SEI size
$k = \Delta f - 4i$
$d = \frac{k}{2}\cos(-2\pi\theta/180); r = \frac{k}{2}\sin(-2\pi\theta/180)$
$S_x = S_y = 0$
$d_x = 0.5R; d_y = 0.5C$
Generate synthetic template
$S((d+s_x)x+(s_y-r)y+d_x,(r+s_y)x+(d-s_x)y+d_y)$
$k = k + i$
Yes
$k \le \Delta f + 4i$
No
STOP

Figure 4.8: Flow chart for generating camera observation (zooming) variable model.

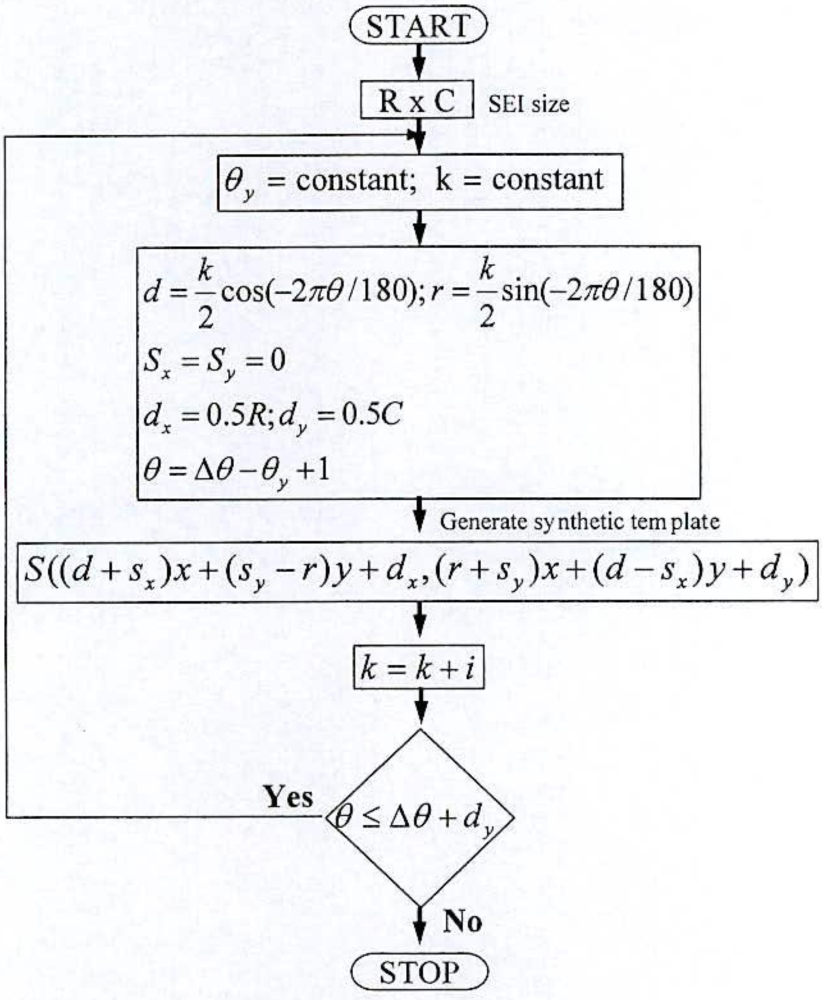

START
R x C
SEI size
θ_y = constant; k = constant
$d=\frac{k}{2}\cos(-2\pi\theta/180); r=\frac{k}{2}\sin(-2\pi\theta/180)$
$S_x=S_y=0$
$d_x=0.5R; d_y=0.5C$
$\theta=\Delta\theta-\theta_y+1$
Generate synthetic template
$S((d+s_x)x+(s_y-r)y+d_x,(r+s_y)x+(d-s_x)y+d_y)$
$k=k+i$
Yes
$\theta\le\Delta\theta+d_y$
No
STOP

Figure 4.9: Flow chart for generating camera observation (slanting position) variable model.

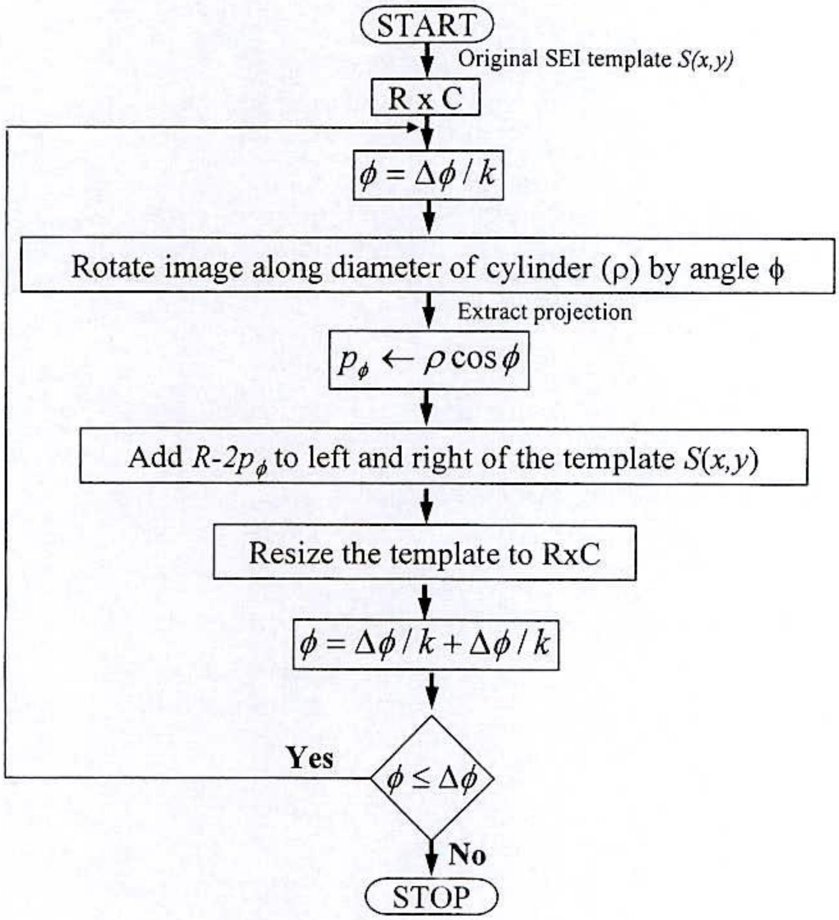

START
Original SEI template $S(x,y)$
R x C
$\phi = \Delta\phi / k$
Rotate image along diameter of cylinder (ρ) by angle ϕ
Extract projection
$p_\phi \leftarrow \rho\cos\phi$
Add $R-2p_\phi$ to left and right of the template $S(x,y)$
Resize the template to RxC
$\phi = \Delta\phi / k + \Delta\phi / k$
$\phi \leq \Delta\phi$
Yes
No
STOP

Figure 4.10: Flow chart for generating human body rotation variable model.

4.2.5. Camera view variations

An action can be viewed by any view of camera. Fig. 4.11 shows the multiple view variations of an action. View 1 shows front view. View 2 shows -45^0 view and view 3 shows $+45^0$ view.

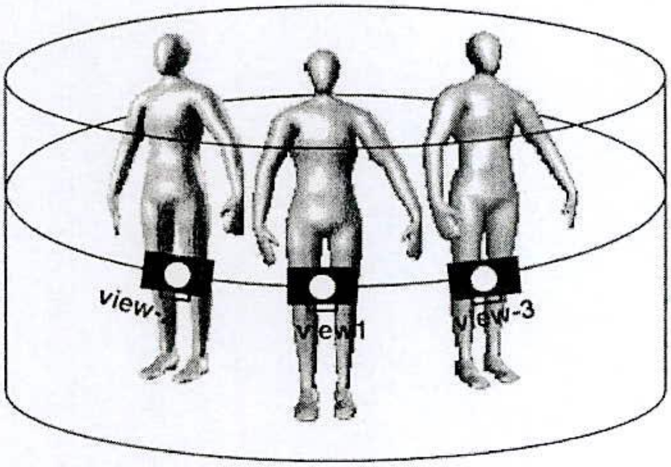


Figure 4.11: Multiple view variations of an action.

CHAPTER 5

Global Motion Descriptions

The action model can be interpreted as a distribution of the motion over the image space in the x and y direction, with the weight $S(x, y)$. Moreover each action model represents a global shape. Global shape of the model with variable intensity is characterized by its local and global motion descriptors.

The adaptable energy templates resembles 2D images with variable intensity due to motion, therefore, model orientation, span, elevation of motions, geometric and orthogonal moments which are the global motion description of the models. Global motion features (GMF) can be integrated by multiple features of action models due to extract more information, and it can be stated as $GMF = (h_i, a_i, Z_{nm}, proj(\lambda_i))$. The symbols are defined in the following subsections.

5.1 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. 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, we consider the relative appropriate position for each action. Hu [15] introduced seven nonlinear functions, h_i, where i=1, 2,...., 7 defined on regular moments using central moments that are translation, scale, and rotation invariant. To achieve the consistent camera observations, we used non-orthogonal features, namely Hu moments, $s_g = \{h_1, h_2, ... h_7\}$ [15] which are slightly modified to extract specific characteristics.

For extracting the Hu-moments, we can use the discrete value of Cartesian moment. The Cartesian moment M_{pq} of order $(p+q)$ for discrete value is given by

$$m_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} x^p y^q S(x,y) \tag{5.1}$$

m_{pq} =Two dimensional Cartesian moment Where M and N are the image dimensions and the monomial product $x^p y^q$ is the basis function.

The zero order moment m_{00} is defined as the total mass (or power) of the image. If this is applied to a binary (i.e. a silhouette) $M \times N$ image of an object, then this is literally a pixel count of the number of pixels comprising the object.

$$m_{00} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} S(x,y) \tag{5.2}$$

The two first order moments are used to find the Centre of Mass (COM) of an image. If this is applied to a binary image and the results are then normalized with respect to the total mass (m_{00}), then the result is the centre co-ordinates of the object. Accordingly, the centre co-ordinates $\overline{x}, \overline{y}$. $\overline{x}$= x-axis centre of mass, $\overline{y}$= y- axis centre of mass are given by:

$$\overline{x} = \frac{m_{10}}{m_{00}}, \qquad \overline{y} = \frac{m_{01}}{m_{00}} \tag{5.3}$$

The COM describes a unique position within the field of view which can then be used to compute the centralized moments of an image.

The definition of a discrete centralized moment as described by Hu is: μ_{pq} = Two dimensional centralized moment

$$\mu_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} (x-\overline{x})^p (y-\overline{y})^q S(x,y) \tag{5.4}$$

This is essentially a translated Cartesian moment, which means that the centralized moments are invariant under translation. To enable invariance to scale, normalized moments μ_{pq} = Two dimensional scale-normalized centralized moments are used, given by:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}, \tag{5.5}$$

Therefore, $$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} = \frac{1}{\mu_{00}^{\gamma}} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} (x-\bar{x})^{p} (y-\bar{y})^{q} S(x,y) \tag{5.6}$$

where $$\gamma = \frac{p+q}{2} + 1, \qquad \forall (p+q) \geq 2 \tag{5.7}$$

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.

Therefore, seven Hu-moments becomes –

$$\begin{aligned}
H_1 &= \eta_{20} + \eta_{02} \\
H_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
H_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
H_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
H_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
H_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
H_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \tag{5.8}$$

Table 5.1: Geometric moments for different actions.

Actions	$\bar{x}$	$\bar{y}$	h_1	h_2	$h3$	h_4	h_5	h_6	h_7
Bowing	46.54	80.89	0.0010	0.00	0.00	0.00	0.00	0.00	0.00
Getting down on the floor	55.18	108.26	0.0010	0.00	0.00	0.00	0.00	0.00	0.00
Lying on the floor	77.11	106.46	0.002	0.00	0.00	0.00	0.00	0.00	0.00
Raising right hand	45.38	86.32	0.0013	0.00	0.00	0.00	0.00	0.00	0.00
Running	66.28	79.44	0.0014	0.00	0.00	0.00	0.00	0.00	0.00
Sitting on a chair	44.13	92.49	0.0014	0.00	0.00	0.00	0.00	0.00	0.00
Sitting on the floor	61.45	114.84	0.0011	0.00	0.00	0.00	0.00	0.00	0.00
Walking forward	91.74	84.91	0.0018	0.00	0.00	0.00	0.00	0.00	0.00
Walking	48.66	77.95	0.0015	0.00	0.00	0.00	0.00	0.00	0.00

5.2 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. 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 invariants. The two-dimensional Zernike moments of an image intensity function $f(\rho,\theta)$ with order n and repetition m is expressed as follows [28].

$$Z_{nm} = \frac{n+1}{\tau}\int_0^{2\tau}\int_0^{1} R_{nm}(\rho)e^{-jm\theta}S(\rho,\theta)\rho d\rho d\theta \tag{5.9}$$

and,

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{(n-|m|)}{2}} \frac{(-1)^s (n-s)!\rho^{n-2s}}{(s)!((n+|m|)/2-s)!((n-|m|)/2)!} \tag{5.10}$$

where,

$$\rho = \frac{\sqrt{(2x-N+1)^2+(N-1-2y)^2}}{N} \tag{5.11}$$

$$\theta = \tan^{-1}(\frac{N-1-2y}{2x-N+1}) \tag{5.12}$$

and $0 \le \rho \le 1$.

For each action model and one given value, we obtain the Zernike moments value. We use the absolute Zernike moment, which is given in (5.13).

$$Z_{nm} = \frac{n+1}{\tau}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} S(x,y)R_{nm}(\rho)\exp(-jm\theta) \tag{5.13}$$

Table 5.2: Zernike moments for different actions.

Actions	**Z_{nm00}**	**Z_{nm10}**	**Z_{nm01}**	**Z_{nm11}**	**Z_{nm20}**	**Z_{nm02}**	**Z_{nm22}**
Bowing	16.53	69.22	16.53	69.22	16.53	69.22	16.53
Getting down on a floor	5.32	35.45	5.32	35.45	5.32	35.45	5.32
Lying on the floor	21.14	9.06	21.14	9.06	21.14	9.06	21.14
Raising right hand	18.20	58.46	18.20	58.46	18.20	58.46	18.20
Running	9.20	78.05	9.20	78.05	9.20	78.05	9.20
Sitting on a chair	25.32	64.86	25.32	64.86	25.32	64.86	25.32
Sitting on the floor	0.92	29.49	0.92	29.49	0.92	29.49	0.92
Walking forward	47.99	21.44	47.99	21.44	47.99	21.44	47.99
Walking	14.50	74.33	14.50	74.33	14.50	74.33	14.50

5.3 Direction of SEI

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 indications 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 variability models.

The covariance matrix is

$$M_c = \begin{pmatrix} \mu'_{20} & \mu'_{11} \\ \mu'_{11} & \mu'_{02} \end{pmatrix} \tag{5.14}$$

The eigenvalue of the covariance matrix is given by

$$\lambda_i = \frac{\mu'_{20} + \mu'_{02}}{2} \pm \frac{\sqrt{4\mu'^{2}_{11} + (\mu'_{20} - \mu'_{02})^2}}{2} \qquad (5.15)$$

where

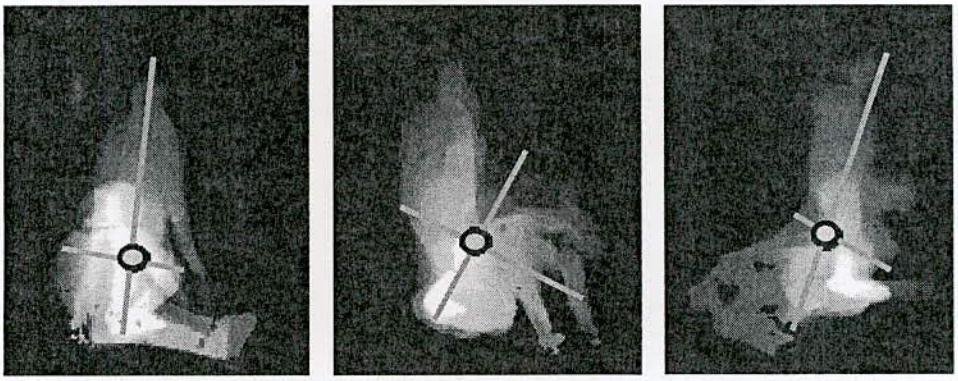
$$\mu_{pq} = \frac{\mu_{pq}}{\mu_{00}} \qquad (5.16)$$

$$\mu_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\left(x - \bar{x}\right)^p \left(y - \bar{y}\right)^q S(x, y) \qquad (5.17)$$

$$\mu_{00} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} S(x, y) \qquad (5.18)$$

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.

Fig 5.1 shows the direction of actions of three different views. The long bar (green) axis indicates the major axis and the short bar (violet) axis indicates the minor axis. The ratio of major axis versus minor axis is different for each action. We have considered the direction of major axis as global features for action models.



(a) (b) (c)

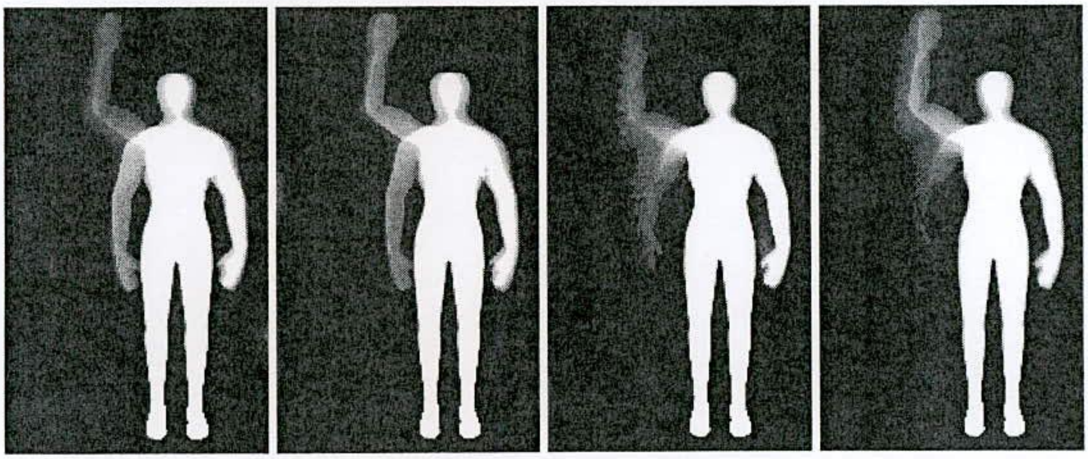
Figure 5.1: Direction of action. (a) Sitting on the floor (b) Getting down the floor (c) Lying on the floor (Direction of action)

Table 5.3: Major and minor axis orientation of actions.

Actions	Major axis	Minor axis	Ecentricity
Bowing	3.62	0.96	0.96
Getting down on a floor	3.25	1.95	0.79
Lying on the floor	17.22	10.37	0.79
Raising right hand	0.34	0.06	0.98
Running	1.44	0.36	0.96
Sitting on a chair	0.88	0.28	0.94
Sitting on the floor	4.04	3.13	0.63
Walking forward	0.20	0.06	0.94
Walking	0.10	0.02	0.96

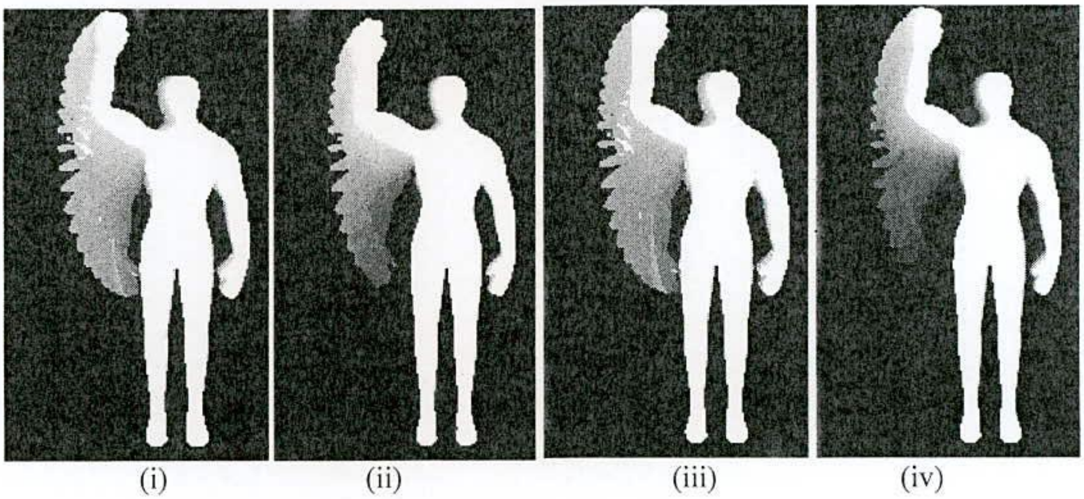
5.4 GM, ZM and direction of SEI and SHI of typical plastic model

The plastic model of SEI and SHI are shown in Fig 5.2 (a) and (b). In each figure (i), (ii), (iii) and (iv) represents images of 1-50 frames,1-60 frames, 20-40 frames, and 20-50 frames respectively. Image size of each figure is 120x240.



(i) (ii) (iii) (iv)

(a)


(i) (ii) (iii) (iv)

(b)

Figure 5.2: (a) SEI for 1-50 frames [i], 1-60 frames [ii], 20-40 frames [iii], and 20-50 frames [iv], (b) SHI for 1-50 frames [i], 1-60 frames [ii], 20-40 frames [iii], and 20-50 frames [iv].

Table 5.4: Geometric moment for SEI and SHI of typical plastic model for 1-50 frames.

Image	$\bar{x}$	$\bar{y}$	h_1	h_2
SEI	91.206200	120.912892	0.0016083349	0.0000015549
SHI	82.022453	105.996147	0.0014303446	0.0000010371

Table 5.5: Geometric moment for SEI and SHI of typical plastic model for 1-60 frames.

Image	$\bar{x}$	$\bar{y}$	h_1	h_2
SEI	91.264580	119.930492	0.0016554454	0.0000016681
SHI	82.762886	107.603023	0.0015341206	0.0000012008

Values of $\bar{x}$, $\bar{y}$, h_1, h_2 of various frames of SEI and SHI are given in Table 5.4 and 5.5.

Table 5.6: Zernike moment for SEI and SHI of typical plastic model for 1-50 frames.

Image	$\mathbf{Z_{nm00}}$	$\mathbf{Z_{nm10}}$	$\mathbf{Z_{nm01}}$	$\mathbf{Z_{nm11}}$	$\mathbf{Z_{nm20}}$	$\mathbf{Z_{nm02}}$	$\mathbf{Z_{nm22}}$
SEI	32.6117	80.5361	32.6117	80.5361	32.6117	80.5361	32.6117
SHI	28.2923	93.2091	28.2923	93.2091	28.2923	93.2091	28.2923

Table 5.7: Zernike moment for SEI and SHI of typical plastic model for 1-60 frames.

Image	$\mathbf{Z_{nm00}}$	$\mathbf{Z_{nm10}}$	$\mathbf{Z_{nm01}}$	$\mathbf{Z_{nm11}}$	$\mathbf{Z_{nm20}}$	$\mathbf{Z_{nm02}}$	$\mathbf{Z_{nm22}}$
SEI	32.9605	77.8945	32.9605	77.8945	32.9605	77.8945	32.9605
SHI	27.6415	84.0635	27.6415	84.0635	27.6415	84.0635	27.6415

Values of different ZM's are shown in Table 5.6 and 5.7.

Table 5.8: Direction of SEI for different actions for 1-50 frames.

Image	Major axis	Minor axis	Ecentricity
SEI	2.903027	1.032802	0.934575
SHI	11.999272	4.921847	0.912005

Table 5.9: Direction of SEI for different actions for 1-60 frames.

Image	Major axis	Minor axis	Ecentricity
SEI	3.434231	1.206759	0.936229
SHI	12.573103	5.132727	0.912879

Parameters like Major and minor axis and ecentricity of the SEI and SHI of 1-50 and 1-60 frames are shown in Table 5.8 and 5.9.

CHAPTER 6

Classification of action

6.1 Different Classifiers

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.

6.2 Bayes classifier

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the popular method, in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers [19]. Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests [20].

An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

6.3 k-NN classifier

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most commonly amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor.

The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. (A common weighting scheme is to give each neighbor a weight of 1/d, where d is the distance to the neighbor. This scheme is a generalization of linear interpolation.)

The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be considered as the training set for the algorithm, though no explicit training step is required. The k-nearest neighbor algorithm is sensitive to the local structure of the data

Nearest neighbor rules in effect compute the decision boundary in an implicit manner. It is also possible to compute the decision boundary itself explicitly, and to do so in an efficient manner so that the computational complexity is a function of the boundary complexity.

6.4 Support Vector Machine (SVM) classifier

Support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The original SVM algorithm was invented by Vladimir Vapnik and the current standard incarnation (soft margin) was proposed by Corinna Cortes and Vladimir Vapnik [21]. The standard SVM takes a set of input data and predictions, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier. Since an SVM is a classifier, then for given a set of training examples, each marked as belongs to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

6.4.1 Learning of SVM classifier

Given a training set of instance-label pairs (x_i, y_i); i = 1,…, l where $x_i \in R^n$ and $y_i \in \{1,-1\}^l$, the support vector machines (SVM) [21] require the solution of the following optimization problem:

$$\min_{w,b,\xi} \quad \frac{1}{2}w^T w + C\sum_{i+1}^{l}\xi_i \tag{6.1}$$

$$\text{subject to} \quad y_i(w^T\phi(x_i)+b) \geq 1-\xi_i \tag{6.2}$$

$$\xi_i \geq 0$$

Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function ϕ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. Though new kernels are being proposed by researchers, beginners may find in SVM books the following four basic kernels:

- linear: $K(x_i, x_j) = x_i^T x_j$
- polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- radial basis function (RBF): $K(x_i, x_j) = \exp\left(-\gamma\left\|\mathrm{x_i} - x_j\right\|^2\right), \gamma > 0$.
- sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$.

Here, γ, r, and d are kernel parameters.

6.4.2 Procedure of SVM classifiers

- Transform data to the format of an SVM package [27].
- Conduct simple scaling on the data
- Consider the RBF kernel $K(x, y) = e^{-\gamma\|x-y\|^2}$
- Use cross validation to find the best parameter C and γ
- Use the best parameter C and γ to train the whole training set
- Test

CHAPTER 7

Experimental Results and Discussions

7.1 Experimental Databases

7.1.1. The KUGDB

The KUGDB [16] 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 are shown in Fig. 7.1. Front view (*v*1), left-side or -45° view (*v*2), and right-side or +45° view (*v*3).

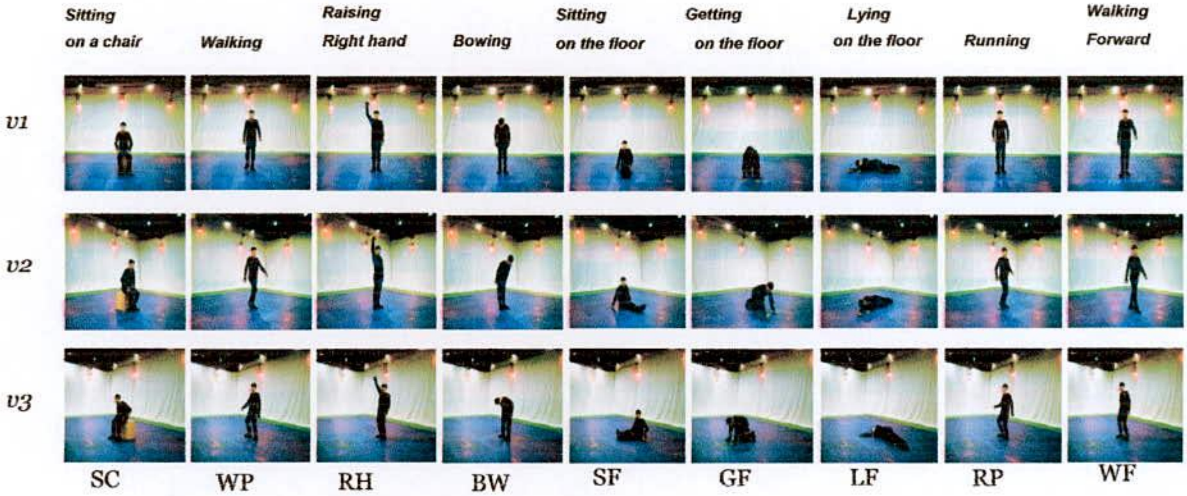


Figure 7.1: Korea University Gesture database (KUGDB) on all specified actions in three different views.

7.1.2 The KTHDB

The KTHDB is one of the largest databases with sequences of human actions taken over different scenarios [17]. 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. 7.2. The image sequences have the spatial resolution of 160×120 pixels and have a length of four seconds in average.

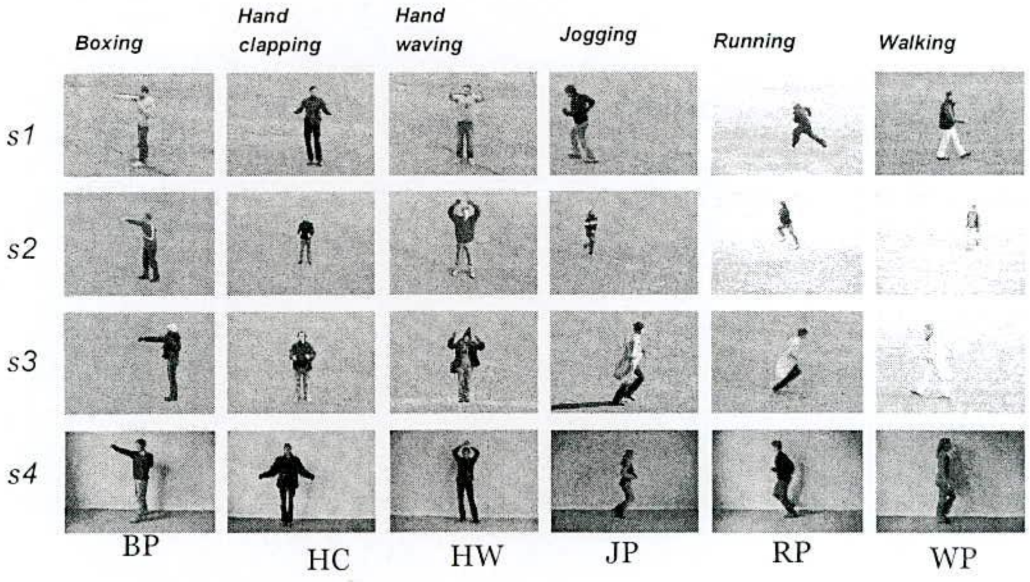
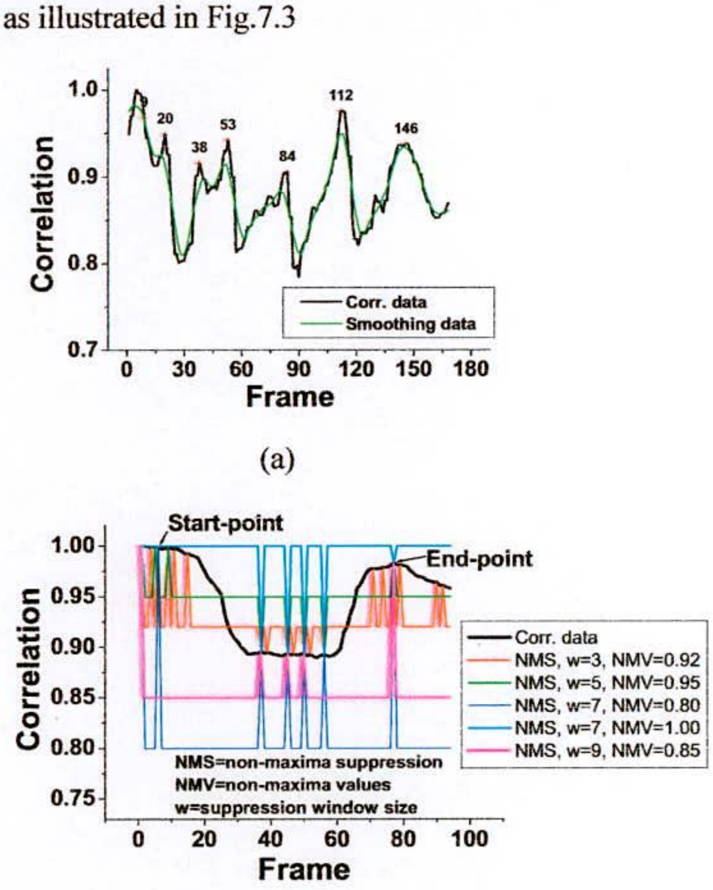


Figure 7.2: Royal Institute of Technology (KTH) human action database on all specified actions in 4 different scenarios.

7.1.3. 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 to 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 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.7.3



(a)

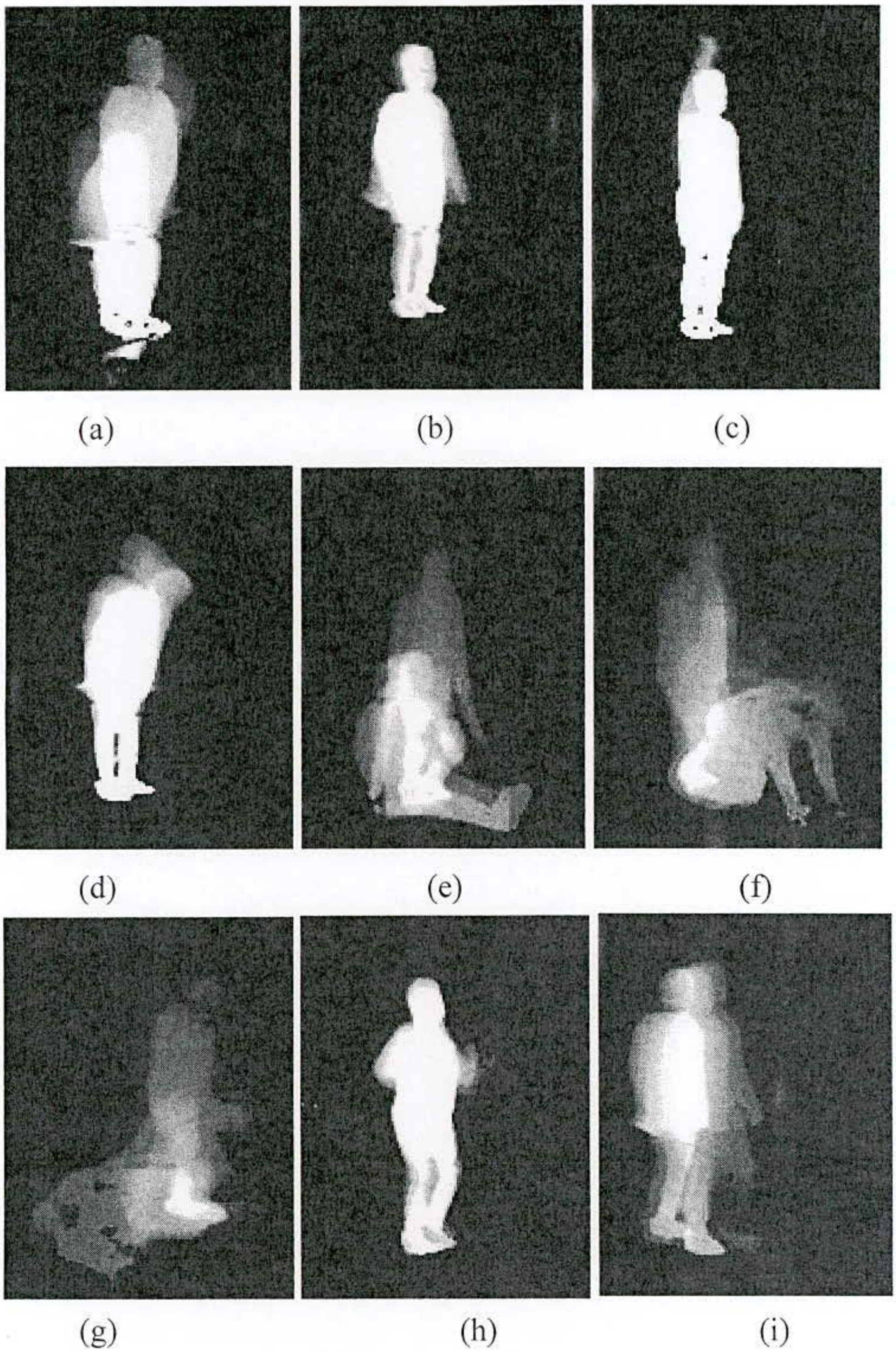
(b)

Figure 7.3: 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 ($t_s = 6$, $t_e = 77$).

7.1.4. Example of SEI and SHI

We consider 9 actions from the KUGDB, where the actions are key actions occurring in daily life. The typical action models are shown in Fig. 7.4. The brighter parts indicate more silhouette energy and the less bright parts indicate less silhouette energy of the action. From the action models, the motion distribution of each action is clearly understood.

Fig. 7.4 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.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 7.4: 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).

7.1.5. Example of Variable SEI

Variable SEI is generated by Using SEI and Variability parameters. From one SEI ,we have produced 40 variable SEI using different variable parameters (anthropometry, speed, Phase, Camera observation).Fig 7.5 shows the SEI of an image sequence (walking at a place) and fig 7.6 to 7.10 show the anthropometric variable SEI.

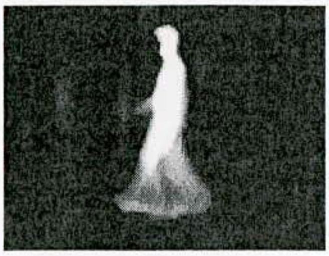


Figure 7.5: SEI of an Image Sequence (Walking at a place)

Anthropometry variable SEI is shown in Fig. 7.6. From the anthropometric variation we get eight different SEIs. They are (a) Higher girth (b) Higher height (c) Smaller girth (d) Smaller height (e) Higher girth and smaller height (f) Smaller girth and smaller height (g) Higher girth and height (h) Smaller girth and height.

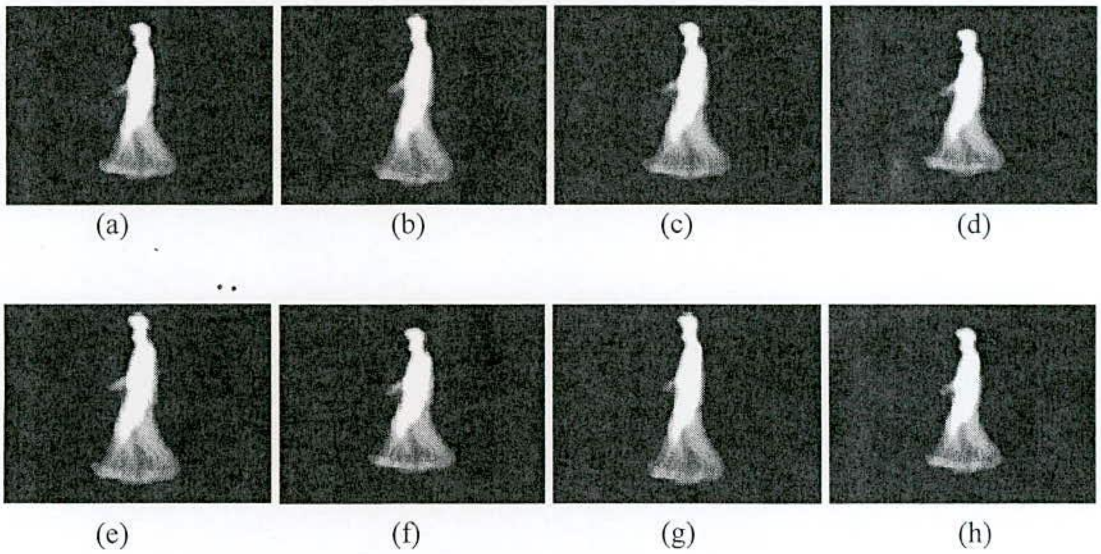
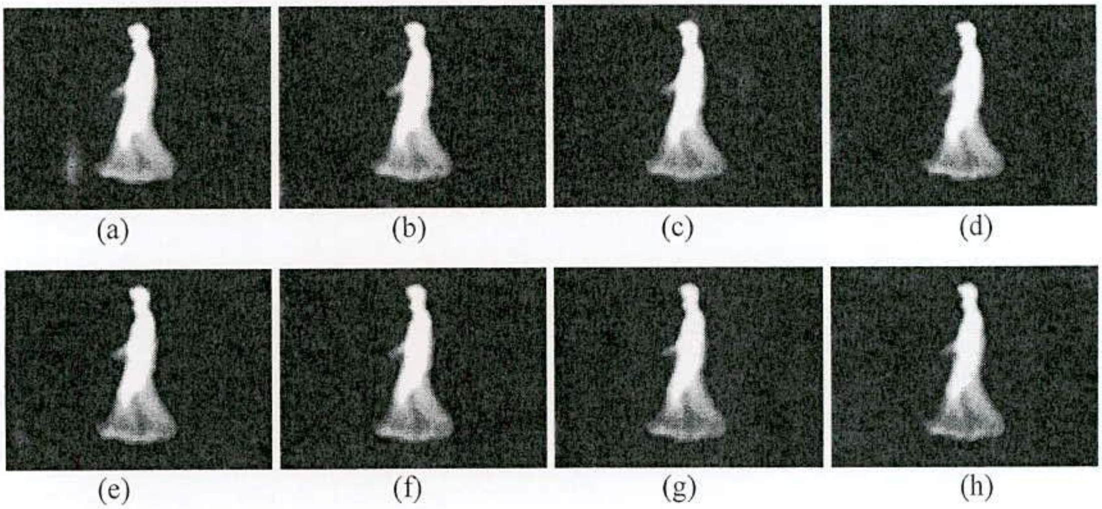


Figure 7.6: Anthropometric Variable SEI. (a) Higher girth, (b) Higher height, (c) Smaller girth, (d) Smaller height, (e) Higher girth and smaller height, (f) Smaller girth and smaller height, (g) Higher girth and height, (h) Smaller girth and height.

Speed variable SEI are shown in Fig. 7.7. We have model speed variable SEI by using eq. 4.4. This does not rigorously follow the speed variation, but it approximates the variation of an action. From the speed variation we get 8 different SEI.

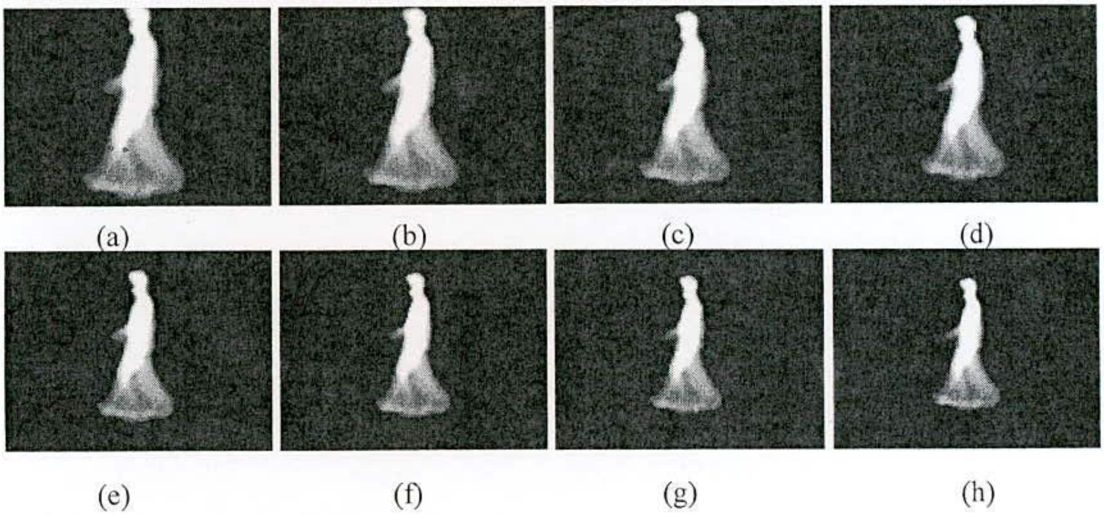


(a) (b) (c) (d)

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

Figure 7.7: Speed Variable SEI. If original SEI has period p, i.e. $p = N$. (a) $p = N$ - 20%N, (b) $p = N$ - 15%N, (c) $p = N$ - 10%N, (d) $p = N$ - 5%N, (e) $p = N$ + 5%N, (f) $p = N$ + 10%N, (g) $p = N$ + 15%N, (h) $p = N$ + 20%N.

Variable SEIs for zooming a camera are shown in Fig. 7.8. This represents person's distance from the camera. The position of person is varied to the frontward or backward of the camera.

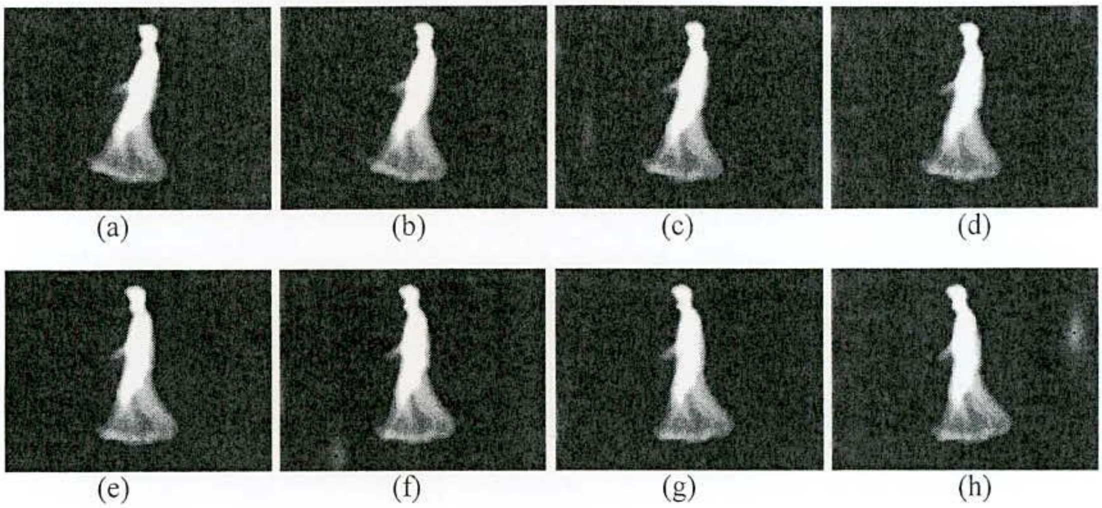


(a) (b) (c) (d)

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

Figure 7.8: Zoom variable SEI if person's distance from the camera is $d = f$. (a) $d = f$ - $4k$, where k is assumed a specific distance, (b) $d = f$ - $3k$, (c) $d = f$ - $2k$, (d) $d = f$ - k, (e) $d = f$+ k, (f) f+ $2k$, (g) f+ $3k$, (h) f+ $4k$.

We get eight variable SEIs from tilting variations that are shown in Fig. 7.9. Tilt angles are varied and shown in figures (a) to (h) accordingly.

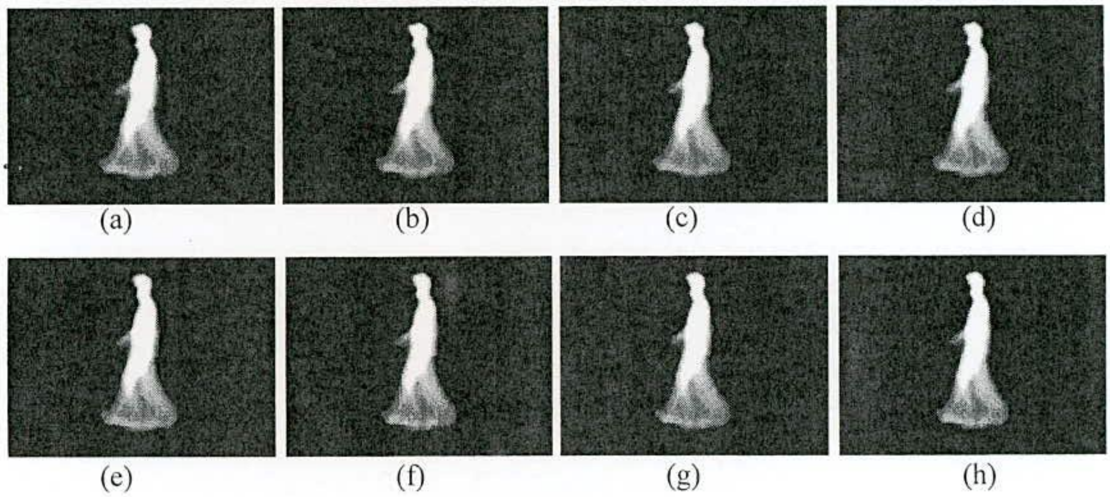


(a) (b) (c) (d)

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

Figure 7.9: Tilt variable SEI. Maximum tilt angle ($\Delta\theta$) from the vertical position to leftward (L_w) or rightward (R_w). (a) $L_w = \Delta\theta$, where $\Delta\phi$ is the maximum range of rotation and k is a constant value, (b) $L_w = \Delta\theta/2$, (c) $L_w = \Delta\theta/3$, (d) $L_w = \Delta\theta/4$, (e) $R_w = \Delta\theta/4$, (f) $R_w = \Delta\theta/3$, (g) $R_w = \Delta\theta/2$, (h) $R_w = \Delta\theta$.

Variable SEI for person's rotation is shown in Fig. 7.10. We get eight variables SEI from person's rotation. The rotation angles are varied and shown in the following figures.



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

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

Figure 7.10: Rotation variable SEI. (a) Rotation angle, $\phi = \Delta\phi/k$, where $\Delta\phi$ is the maximum range of rotation and k is a constant value, (b) $\phi = 2\Delta\phi/k$, (c) $\phi = 3\Delta\phi/k$, (d) $\phi = 4\Delta\phi/k$, (e) $\phi = 5\Delta\phi/k$, (f) $\phi = 6\Delta\phi/k$, (g) $\phi = 7\Delta\phi/k$, (h) $\phi = 8\Delta\phi/k$.

7.1.6 Classification Result

The accuracy or correct recognition rate (CRR) was defined by (7.1). The expression for CRR can be written as

$$CRR = \frac{N_c}{N_a} \times 100\% \qquad (7.1)$$

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

Table 7.1: CRR of each action and view of KUGDB

View	SC	WP	RH	BW	SF	GF	LF	RF	WF
v1	.98	0.57	0.86	1.0	0.71	0.57	1.00	0.57	0.86
v2	.85	0.44	1.00	0.86	0.86	0.86	1.00	0.86	0.86
v3	.97	0.86	1.00	1.0	0.86	1.0	1.00	0.57	0.72
vA	.94	0.57	0.94	0.99	0.76	0.76	0.90	0.76	0.76

Table 7.1 shows the action recognition results of KUGDB using SVM classifiers where we use the global motions for each view. Nine subjects, 9 actions, and 4 views variation were used for testing (vA represents arbitrary view). As can be seen, there is a clear separation among different kinds of actions.

Boxing
Hand clapping
Hand waving
Jogging
Running
Walking

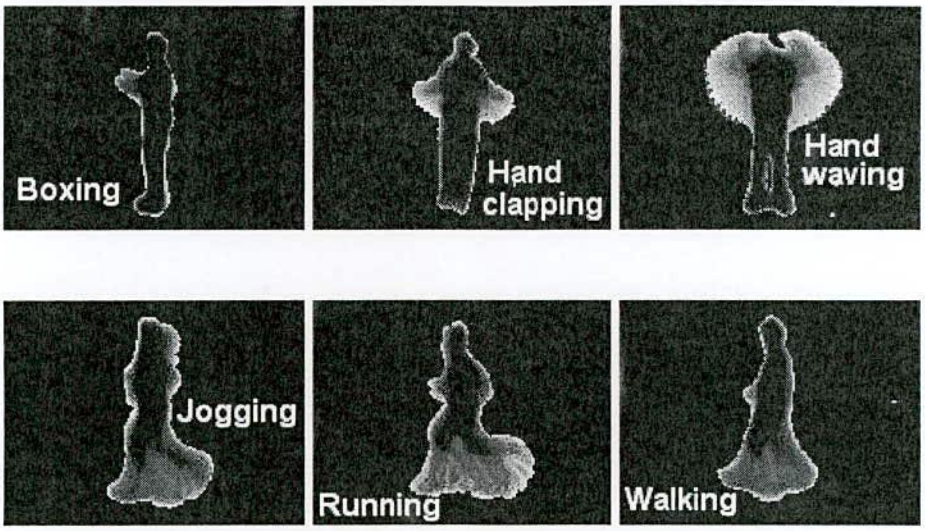


Figure 7.11: Human action models and corresponding motion distribution (KTHDB). First and second rows show the SEI. Third and fourth rows show the motion distribution of corresponding SEI.

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

Table 7.2: CRR of each action and view of KTHDB

Scen.	**BP**	**HC**	**HW**	**JP**	**RP**	**WP**
s1	0.98	0.97	0.97	0.91	0.74	0.82
s2	0.98	0.98	0.94	0.95	0.61	0.78
s3	1.00	0.96	0.97	0.74	0.81	0.81
s4	0.95	0.94	0.93	0.81	0.67	0.80

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 7.2 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 (KUGDB). 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 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.

Table 7.3: Comparison of Action Recognition.

Method		**Recognition accuracy**	**Scenarios**
Niebles et al.	[8]	81.50	All scenarios
Dollár et al.	[7]	81.17	All scenarios
Jiang et al.	[9]	84.43	All scenarios
Schüldt et al.	[5]	71.72	All scenarios
Ke et al.	[6]	62.96	All scenarios
Our method		87.50	All scenarios

Our works are compared 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 [5], [6], [7], [8], and [9] 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 7.3. Compared to the mentioned researches, our approach yields the best recognition results.

CHAPTER 8

Conclusion

We proposed a novel method for human actions model using silhouette energy image (SEI) and silhouette history image (SHI) with variable silhouette energy image. 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. Here we have proposed different variations such as, anthropometric variation, phase variation, speed variation, phase variation, camera view variation etc. Using an advanced human-machine interface these variations provided a more natural and robust environment for human action recognition. From the variable SEI or variable action mode, global motion properties are extracted. We recognized different daily human actions successfully in the indoor environment as well as in the outer environment.

In this thesis we have considered some assumptions that make the recognition of human actions a challenging task. We recognized human actions in individual and arbitrary views. The action recognition rate might not be extremely higher but it was shown that we recognized actions from any view rather than a set view. Here we have used two databases namely, KUGDB and KTHDB. We mainly used multi-class SVM for each action.

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. We did not use a huge number of image sequences for learning incase of lack of image sequences, since a few number of sequences were adequate for modeling and recognition actions. Moreover, by detection of period or duration of an action from image sequence, we need not considered all frames in the image sequences.

Despite of robustness we faced the problem of recognition in actions. This happened when two or more actions had similar body silhouette maps and global motion variation was low over long image frames. Due to 2-D representation of human actions the current limitation is the direction of action, for example, sitting on a chair or standing from a chair showed the same recognition result. Slow running and walking could not be distinguished. Our future work will include the precise detection and recognition of action in more complicated situations.

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] C. Schüldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," *Proc. IEEE Conf. ICPR*, vol. 3, pp. 32-36, 2004.

[6] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," *Proc. Int'l Conf. Computer Vision*, pp. 166-173, 2005.

[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] 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.

[9] 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.

[10] O. Masoud and N. Papanikolopoulos, "Recognizing Human Activities," *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, pp. 157-162, July 2003.

[11] 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

[12] 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.

[13] Y. Sheikh and M. Shah, "Exploring the Space of a Human Action*," Proc. IEEE Int'l Conf. Computer Vision,* pp. 144- 149, October 2005.

[14] A.Yilmaz and M. Shah,"Matching Actions in Presence of Camera Motion,"*Computer Vision and Image Understanding*, vol. 104, no. 2,pp221-231, Nov.2006

[15] M-K. Hu., "Visual Pattern Recognition by Moment Invariants," *IRE Trans. On Information Theory*, IT-8, pp. 179-187, 1962

[16] The KU Gesture Database *http://gesturedb.korea.ac.kr/.*

[17] The KTH Database, *http://www.nada.kth.se/cvap/actions/.*

[18] M. Ahmad and S.-W. Lee "Human Action Recognition using Shape and CLG motion flow from Multi View image sequence", *Pattern Recognition*, vol. 41,

pp.2237-2252, 2008.

[19] H. Zhang "The Optimality of Naive Bayes". *FLAIRS2004 conference*.

[20] R. Caruana, and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms". *Proceedings of the 23rd international conference on Machine learning,* 2006

[21] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning, 20,* 1995.

[22] S. Seitz, C. Dyer, View invariant analysis of cyclic motion, *Int. J. Comput. Vision* vol. 25, pp. 231–251, 1997.

[23] C. Rao and M. Shah, "View-invariance in action recognition," *Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, pp. 316–323, December 2001.

[24] V. Parameswaran, R. Chellappa, "View invariants for human action recognition," *Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 613–619 , 2003.

[25] M. Ahmad, I. Parvin and S. Lee," Silhouette History and energy image Information for Human Movement Recognition," *Journal of Multimedia, Academy Publisher*, pp 12-21, 2010.

[26] I. Parvin and M. Ahmad, "Global and local motion descriptions for human action recognition," *IEEE Int'l Conf. on Electrical and Computer Engineering (ICECE)*, pp. 678-681, Dec. 2010.

[27] C. W. Hsu, C. C. Chang and C. J. Lin,"A Practical guide to Support Vector Classification",2009.

[28] 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.

PUBLICATIONS

[1] M. Ahmad, I. Parvin and S. Lee, "Silhouette History and energy image Information for Human Movement Recognition," *Journal of Multimedia, Academy Publisher*, pp 12-21, 2010.

[2] I. Parvin and M. Ahmad, "Global and local motion descriptions for human action recognition," *IEEE Int'l Conf. on Electrical and Computer Engineering (ICECE)*, pp. 678-681, Dec. 2010.