August, 2012
Ph.D Dissertation

Silhouette Edge-based Log-polar Descriptor for Human Action Representation and Recognition

Graduate School of Chosun University

Department of Computer Engineering

Wilfred Onyango Odoyo

Silhouette Edge-based Log-polar Descriptor for Human Action Representation and Recognition

실루엣 윤곽선 기반의 Log-polar Descriptor를 이용한 동작 인식

24th, August 2012

Graduate School of Chosun University

Department of Computer Engineering

Wilfred Onyango Odoyo

Silhouette Edge-based Log-polar Descriptor for Human Action Representation and Recognition

Supervisor: Professor Beom-joon Cho

This Dissertation is submitted to the Graduate School of Chosun University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Engineering

April, 2012

Graduate School of Chosun University

Department of Computer Engineering

Wilfred Onyango Odoyo

This submitted Doctoral Dissertation of Wilfred Onyango Odoyo is accepted

Committee Chair	Professor, Chosun University	Yun-Bae Lee
Committee Member	Professor, Chosun University	Yong-Geun Bae
Committee Member	Professor, Chosun University	Sang-Woong Lee
Committee Member	Professor, Chosun University	In-Kyu Moon
Committee Member	Professor, Chosun University	Beom-Joon Cho

June, 2012

Graduate School of Chosun University

Contents i

List of Figures

List of Tables

Symbols and Initials

OGD	-	Own Generated Database
KTH	-	Kungliga Tekniska högskolan Action Database
HD	-	Hausdorff Distance
Log-polar	-	Log-polar transformations
MIE	-	Motion Energy Images
MHI	-	Motion History Images
MLD	-	Moving Light Displays
UCF	-	University of Central Florida
2D	-	Two dimensional
3D	-	Three dimensional
LPT	-	Log-polar transformation
ROI	-	Region of Interest
SVM	-	Support vector machine
H(P)	-	Shannon's information entropy
LOO	-	Leave-one-out
DB	-	Database
$E_{\max}$	-	Maximum entropy
$E_{\min}$	-	Minimum entropy
T	-	Threshold

$H(A,B)$	-	Hausdorff distance between A and B
$H(B,A)$	-	Hausdorff distance between B and A
d_{trans}	-	Distance transformation function
SIFT	-	Scale-invariant feature transform
GLOH	-	Gradient location and orientation histogram
PCA	-	Principle component analysis

요 약

실루엣 윤곽선 기반의 Log-polar Descriptor를 이용한 동작 인식

By Wilfred Onyango Odoyo
Advisor : Prof Beom-joon Cho Ph. D
Department of Computer Engineering
Graduate School of Chosun University

인간 동작 인식은 수많은 영상처리 기법과 컴퓨터 비전에 있어서 수요가 지속되는 응용 분야로 인간이 지능을 가진 기계와 상호작용하는 방법을 향상시키는데 있어 필수적인 연구가 되어왔다.

본 논문에서는 동영상에서 인간 신체의 움직임을 추적하고 실루엣을 추출하며 신체 부분의 동작을 결합하여 동영상 내의 동작을 인식하는 방법을 제시한다. 인간 동작이 내포된 동영상은 동작을 표현하는 자세(posture)가 포함된 프레임으로 나눌 수 있다. 각각의 동작을 대표할 수 있는 주요 프레임을 검출하게 되면 적은 데이터 처리량으로 빠른 동작 인식이 가능하므로, 본 논문에서는 주요 프레임 검출을 위해 정보 엔트로피를 사용한다. 동영상 시퀀스의 한 동작에 해당하는 프레임들 간의 정보 엔트로피를 비교하여 임계값 이상의 변화량을 갖는 프레임은 자세 변화가 크므로 동작 인식을 위해 주요 프레임으로 선택된다.

그리고 특징 추출을 위해 선택된 프레임들의 실루엣 경계인 에지를 검출한다. 에지는 중요한 형태 정보를 보존하고 있으므로 인간 동작을 모델화하고 표현하는데 사용될 수 있다. 선택한 프레임들의 실루엣 에지의 결합은 각 동작을 구별할 수 있는 독특한 패턴의 영상을 만들어 내며, 이러한 패턴 영상은 결합 형태를 통해 동작

들 간의 유사성과 비유사성을 보여준다. 그러나 에지 결합 영상은 동일한 동작임에도 실루엣의 크기 및 회전으로 인해 동작 인식의 어려움과 오인식의 문제가 존재한다. 따라서 동일한 동작이 서로 다른 동작으로 분류되지 않도록 크기와 회전에 강인한 기법이 요구되며, 본 논문에서는 이를 해결하기 위해 에지 결합 영상을 거리 변환 후 log-polar로 변환하는 방식을 제안한다.

거리 변환은 영상의 각 픽셀에 가장 근접한 검은 픽셀과의 거리를 할당하는 방법으로 픽셀들이 영상의 패턴을 분명히 표현하므로 두 영상을 구분하는 특징으로 사용된다. log-polar 변환은 각 동작의 에지 결합 영상을 직교 좌표 공간에서 극좌표 공간으로 이동하여 반경과 각도를 이용하므로 형태 크기에 관계없이 동일한 동작들은 유사한 분포를 나타내는 특징이 있다. 이러한 log-polar 변환을 통해 획득한 영상들은 동작 인식을 위해 원형 정합(template matching)을 사용하며 유사성의 측도로 Hausdorff 거리를 사용한다.

본 논문에서는 실험을 위해 KTH 데이터베이스가 제공하는 6가지 동작과 Weizmann 데이터베이스의 8가지 동작, 그리고 OGD 데이터베이스의 8가지 동작을 사용하였으며, 제안한 방법을 적용한 log-polar 변환과 Hausdorff 거리를 적용한 결과, 동일한 동작 간의 거리는 0이거나 아주 작은 값이었으며, 동작의 유사성이 클수록 거리 측정값은 작았고, 서로 다른 동작들은 큰 거리값을 나타냈다. Weizmann 데이터베이스와 OGD 데이터베이스 간의 인간 동작의 유사성을 측정하기 위해 교차 거리를 측정하였으며, 실험 결과는 본 논문에서 제안한 방법이 동작 인식에 있어 우수함을 보여주고 있다.

Chapter 1

Introduction

1.1 Research motivation

Computer vision is a very dynamic field just as the many challenges that computer vision is supposed to provide solutions for. Object recognition, motion analysis in various scenes, biometrics, medical images analysis, are just but to mention a few that depend on the technology provided by computer vision methods. Human action recognition is one such problem that require amicable solutions to its varied applications. There are as many applications to human action recognition just as there exist numerous actions that the human body can express with its poses or movements. Human eyes can detect almost instantly what action has been performed by an individual without making assumptions to it. This is possible even in very complicated scenes with multiple persons.

The call for computer vision in human action recognition is not to outdo the human eye but to enhance security in places where necessary, indexing and retrieving video files with a labeled action in it with ease, give chronicle of events in case of crime scene evidence, analyze and ascertain that the action took place. In this dissertation we try to model typical human activities for recognition in videos. These include running, walking, one-hand wave, two-hands wave, jump-forward, jump-jack, bending, spot- jump and jogging. The actions seem very easy to identify through human eyes but computer vision faces a number of challenges that deter machines from perfect recognition. These challenges could be externally infused or due to the varied postures exhibited for one action by different actors / agents. Recording environment and capture settings, intra- and inter-personal differences, illumination, self occlusion, and scene clutteredness, and noise compose some of the challenges needed to overcome.

Some of the researched and developed techniques have tried to solve the afore-mentioned problems with quite acceptable degree of success. Figures in 1.1(a), (b), and (c) show some examples of typical actions from three different databases of KTH database, Weizmann database, and database generated by AI&PR Lab of Chosun University for this research work.

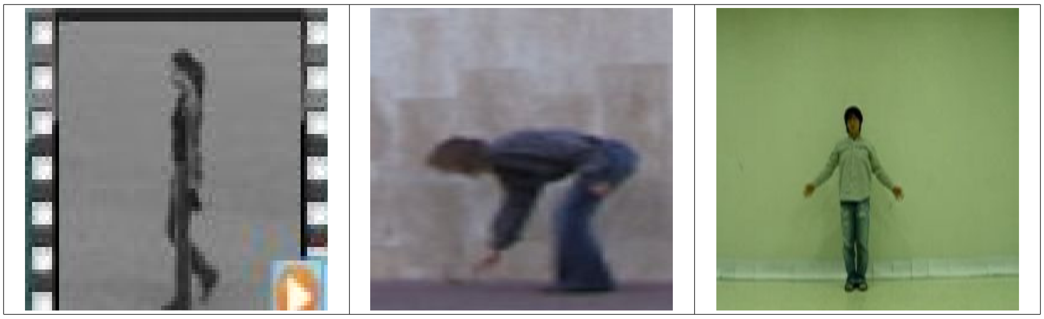


Figure 1.1 Snap shots of video captures of typical actions of walk, bend and two-hand wave from three different databases. (a) KTH database (b) Weizmann human action database (c) Our own-lab generated action database.

1.2 Research context

Vision-based human action recognition methods can be summed up as combinations of processes that involve feature extraction followed by classification of the image representations. Main researched approaches circle around silhouettes of human body [1], pose estimation [2], and spatial - temporal domains [3] of the agent. Some of the methods which base their techniques on the mentioned approaches are the use of temporal templates [4], angles of inclination [5], global motion features [6], and exemplar poses [7]. They all interpret the human body depending on the region of interest for feature extraction and representation. The levels of interpretation could be at action primitive level, action level , or activity level of interpretation. Action primitive concerns the movement of the body limbs (hands and legs). A consecutive movement by the action primitive constitute action level while a series of actions form the activity as we see it. This activity is what humans interpret as running, walking, or any other activity that we seek to label. The methods differ from each other depending on the feature selected for representation and discrimination of different activities. Some methods extract features from the whole body while others consider the time when the action was executed. Angles of inclination of either the torso or the other parts of the body have been considered, too. The shape an exemplar takes after modeling is also considered by other methods among others. Figure 1.2 (a, b, c) are some shapes of models or extracted features for similarity measurements by some of the methods mentioned above.

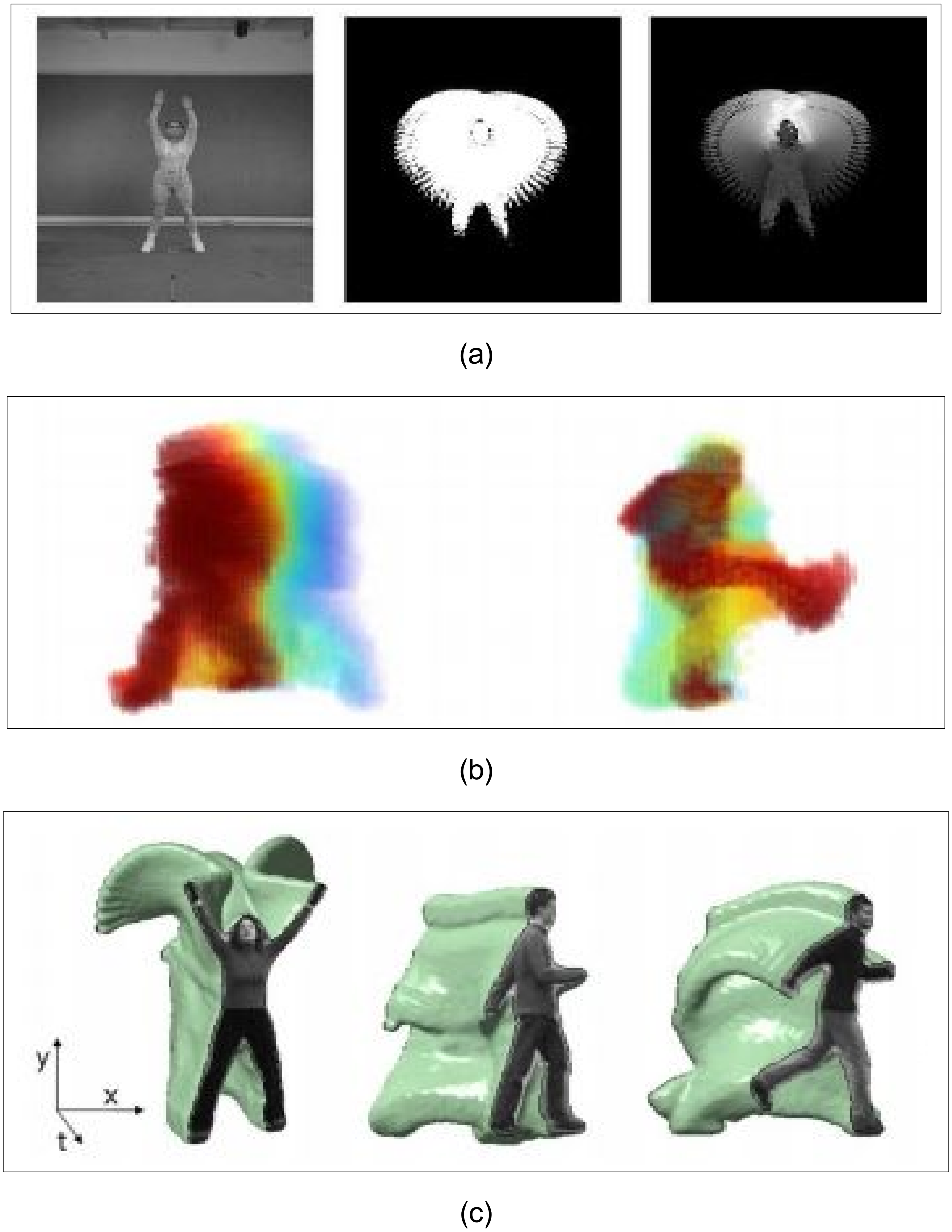


Figure 1.2 Some feature representations presented by different methods. (a) motion energy images (MEI) and motion history images (MHI) [4]; (b) motion history volumes (MHV) [8]; (c) space-time shapes [9].

1.3 Contributions

Recognition method we are putting forward here resembles in operation the one carried out by vision processing of human beings. Humans can recognize poses or activities mostly by recalling from memories similar examples of the past events. The neural system trains and stores actions according to the universal labeling of those activities. The new approach that we are proposing starts with the extraction of silhouettes from a given video sequence. Key postures are selected from a sequence of extracted silhouettes through information entropy. The silhouette boundary edges are extracted from those plausible series of poses selected for the key frame list and then modeled by stacking the edges together to form particular shape / pattern depicting a known activity. We get the log-polar form of these patterns. These shapes become the gallery images that shall be used in discriminating the incoming probe image during classification stage. The probe images are also modeled as per the explained procedure above. Similarity measure is through a distance measure between the processed probe and the modeled gallery images based on some threshold. In our case, Hausdorff Distance measure has been implemented. Euclidean distance measure has been implemented for a comparison to strengthen the belief that our modeled exemplars are fine enough to distinct the different actions.

We identify the main contribution as the collection of spatial-temporal edge features belonging to the detected human silhouette summed up to model and represent the actions in question. The silhouette edges extracted is presented in log-polar image form as our descriptor to the actions.

Silhouette edge-based log-polar descriptor used comes with merits worth mentioning. First, it provides for easy representation through the use of edge features. Next, the computation cost is immensely reduced due to dta reduction in terms of the amount of data used. Third, the use of log-polar transformation solves the scale-invariance problem posed due to different sizes of actors. Another advantage is the salient representation of action packed region of interest (ROI) the method exhibits. The method proposed here can also be easily embedded into existing systems for its applications. Last but not least, with the advancement made in image preprocessing techniques, the feature extraction and representation will be highly accurate.

As will be seen later, our method suffers occlusion problem and could increase computation cost if it is to be addressed. The method proposed does not keep history images as with MHI images, which could be very helpful in monitoring the sequent of events. Complex backgrounds with multiple activities is an uphill task for our method to extract target objects and model.

1.3.1 Inside the proposed method

i) The general assumption here is that all our subjects are human beings and the system is only designed to model human figures from a video. Occlusion is a problem, therefore if the human body is not fully detected and represented, this will affect recognition.

ii) A video from a single camera has been used here and 2-dimension images extracted in form of silhouettes.

iii) The action in the video is modeled from a sequence of images extracted in a consecutive manner to capture every moment and movement observed. Entropy is used to reduce redundancy of similar frames being selected.
iv) The models we construct are of either frontal or sideway views of the human body. This makes our method rigid and can only be used in application specific cases. The method for constructing the descriptor can widely be applicable only if a case is isolated and trained to conform with the recognition process put forward in this thesis.

The general framework of the proposed method is shown in Figure 4.1. Each part of the framework has either been covered as a separate chapter or as subsections of other chapters.

The rest of this dissertation is organized in the following way. The second chapter is filled with a range of related works that have been explored and ongoing in the area of human action recognition. We present three main approaches widely used focusing on the Motion History Images (MHI) and Motion Energy Images (MEI) which our proposed method borrows heavily from. In chapter three, we cover the theory of Information Entropy and its application within our framework. Other important image processing methods like mathematical morphology, distance transformation, and log-polar transformations have been given attention, too. For recognition purposes, details of Hausdorff Distance measure, Chamfer Matching, and Euclidean distance measure have been explained in the same chapter. The step-by-step procedure of *Silhouette edge-based log-polar descriptor*, our proposed method, is dealt with in chapter four. Expect to get the conclusive work from preprocessing, feature extraction, feature representation, and recognition using the developed descriptor in this chapter. The same chapter also reveals the pseudo-algorithm through which our descriptor has been derived. Experimental results and analysis occupy most part of chapter five. In addition to the KTH and Weizmann human action databases, our own-generated database (OGD) have been explained and experiments validating our proposed method shown. We conclude our work in chapter six and give further direction our research will take in the future.

Chapter 2

Related Works

This chapter is dedicated to the works that have been published in human action recognition. It contains the vision-based methods for human action feature extraction, representation, segmentation, and recognition.

2.1 Introduction

Due to the many uprising applications of activity recognition in various fields, human action recognition has attracted a lot of research interest in computer vision. The race in this is to better the human action agent's feature extraction, represent these features in a robust manner that can be adapted in more than one application, and lastly to improve the performance of recognition machines. We have the same goal in this work of seeking to represent and recognize typical human actions like running, bending, waving, jumping, and other common activities. The research work we present try to classify human actions based on the observation of the whole human body motions. As a fact, most literature reviewed present methods based on spatial and temporal dimensions. Spatial domain concentrates where the action happened and is based on the global image features, parametric image features, and local image features. Temporal is all about the time when an action took place. Here recognition is based on global temporal signatures, grammatical models and sparse and unstructured observations [10].

In this work we combine the two aspects of the domains stated above to come up with a new descriptor for representing actions for recognition, and is robust in nature. The silhouette boundary-edges extracted contain both the global and local features from the spatial domain. The edges extracted are also temporal by nature of being stacked to represent the images from video sequences key frame list. The various representations presented in most of the reviews have the aim of providing salient image features that can properly discriminate one action from the other based on the pose and motion detected. They also differ on how much information can be extracted from the actor / agent in a video.

There are three main approaches used in action recognition. These are;

i) Template matching

ii) State-space approaches, and

iii) Semantic description of human behavior.

Some other research work might categorize them differently as image models, body models and sparse features, but all are similar in the fundamental operations. The main difference in the three mentioned approaches is the environment in which they are to be applied, the nature of the activity in a video sequence, the speed and the external effects on the video. Below, we explain the three approaches.

2.2 Model representations

As mentioned above, the three widely published methods of representation include template matching, state-space approaches, and semantic description. Because our work is deeply rooted in the usage of global image to extract features for representation, we will dwell more on the variants of image models from globally extracted features as the foundation of our work.

2.2.1 Body models

This method of representation involves extraction of human body features from a sequence of video frames. It is biologically plausible approach and finds support from psychophysical work on visual interpretation of biological motion [11]. Here a new method, moving light displays (MLD) is explained [12]. This work inspired other similar ones based on landmark points on the human body. The argument here is that human beings have the capability of recognizing actions by observing the light movements mapped on the specific points of a human body. The same work gave rise to arguments of 3D and 2D models for action recognition. 3D models are not in our framework and is reserved for future work. Direct recognition from 2D images modeled from human silhouette edges constitute our work. [13] reckons that regardless of whether a method uses 2D or 3D models for human action recognition, locating different body parts and estimating parametric body models is the big matter in action recognition and is still not a trivial matter in the field of computer vision. MLD, shown in Figure 2.1 are models of two different movements.

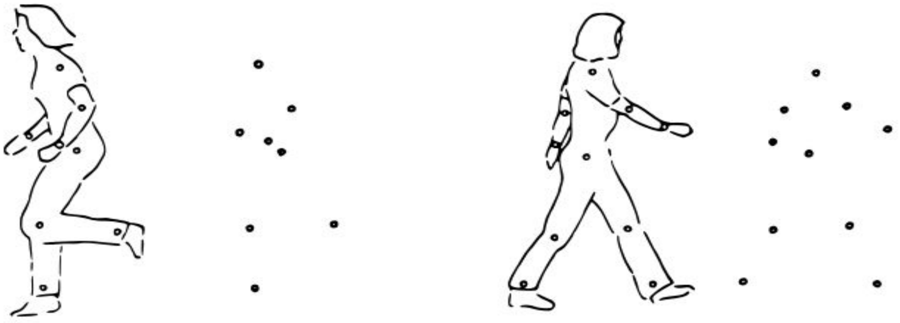


Figure 2.1 Illustration of moving light displays (MLD) [12] to model the body movements

2.2.2 Sparse features

These are based on sparse image features from detected interest regions and organized into a spatial bag-of-features. The local representation of action decompose an image or video into smaller interest regions and give them each a description of their own. The actions are recognized based on the scattered-collected features. The features are mostly detected in corners or blob like structures and labeled with a vocabulary of a bag-of-words. Bag-of-words refers to a histogram that count how many times the vocabulary describing the feature occurred within an image. This method has gained popularity and has been researched with so many variants of it as in [3 14 15 16 17].

There has been successes attributed to this method. It is straightforward and easily applicable on difficult scenes, and they are less affected by occlusion compared to global features. They provide view invariance to affine transformations. However, scenes with multiple persons is still a challenge among other issues.

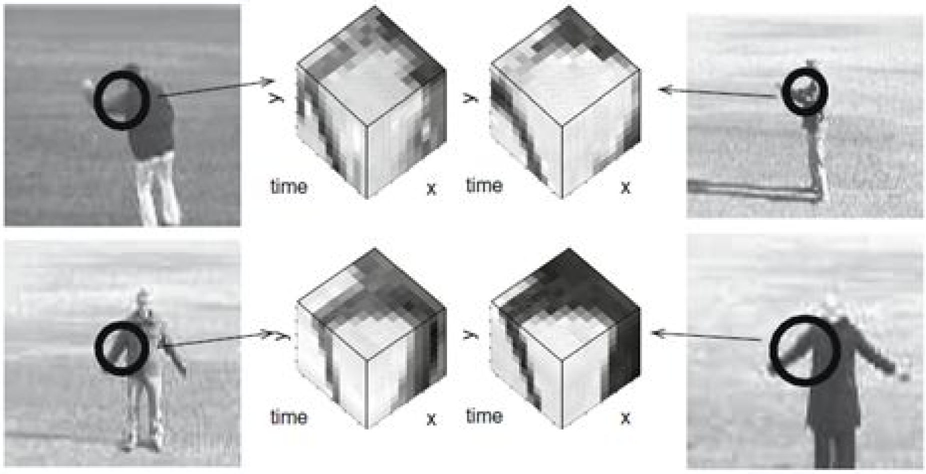


Figure 2.2 Extraction of space-time cuboids at interest points from similar actions performed by different persons [31]

The local space-time approach mentioned above applies motion recognition without going through image segmentation step. Detected neighborhoods known as local motion events based on space-time domain provide information for discrimination. Figure 2.1 above, taken from the work [31], shows local space-time neighborhoods for corresponding space-time points I image sequences of 'boxing' and 'two-hands waving'. Examining the cubes in the middle, the correspondence in the neighborhoods can be seen as similar even though the variations in the actors clothing and viewing distance is obvious. Through the patterns formed, a clear difference is revealed even between two different actions.

2.2.3 Template / image models

Template is a representation of a particular action or pattern as exhibited by the features extracted. This third type of model can be categorized in the global representation where by the features are extracted and modeled usually as a whole. This requires that the object / person in the video is first localized by using background subtraction, or image differencing. The descriptor is as a result of taking the image wholesomely as a region of interest and representing it. The models from this representation contain much information about the image but they require extra caution in initialization for salient background removal methods.

Most globally represented templates are derived from the human silhouettes optical flow, or edges. In this dissertation we apply the latter to construct our newly proposed descriptor. We note here that even though the global approach was used in extracting the silhouette edges, the descriptor formed captures local features owing to the fact that the intensity is higher in particular areas where the motion or movements most frequently observed. For example, in a visual observation, a one-hand wave will produce a darker or thicker pixel values on the right or top-left corner of the human silhouette depending on which arm raised. Many researchers have pursued this type of representation to come up with robust descriptors. We will discuss some of them and their variants that have come up so far in line with our work. Most notably are the Motion Energy Images (MEI) [18], and Motion History Images (MHI) [19].

Our approach is considered a variant of the many view-based temporal template methods that have been researched and implemented in various applications. As introduced above, we will consider a widely used and improved method called Motion History Images (MHI). Similar to our approach is also an earlier version called Motion Energy Images (MEI) [18], a template method which first converts image sequence into recognizable static shape pattern. At the time of matching, the modeled static shape pattern from an input sequence video is compared to the prestored action prototypes for recognition [20].

2.2.3.1 Motion Energy Images (MEI)

This can be defined as the cumulative binary motion image that can describe where a motion is in a video sequence. Suggested by A. Bobick and J. Davis [18, 19], MEI is a view-based approach to the representation of an action. Their method computes MEI to grossly describe the spatial distribution of motion energy for a given view of a given action. Main argument here is to establish and model where the motion occurred as opposed to how the motion happened.

We borrow heavily from the idea that the selected key frames from the video sequence are accumulated to form one frame comprising the action leading us to the final decision that an activity took place. This final shape pattern depicts that an action actually occurred and the viewing condition can also be seen, be it frontal or side. Bobick et. al. [18] described the motion pattern by first constructing the MEI for each training sequence. Using local, gradient technique [21], optical flow field [22] between each pair of frames is computed to get a vector image I(x,y) for a sequential pair. The motion energy image is computed as;

$$MEI(x,y) = \sum_{i}^{T} I_i(x,y) \tag{2.1}$$

While in our work we have used entropy information to select key frames, [23] used binary thresholding of the sum of squared difference between each frame and the first one.

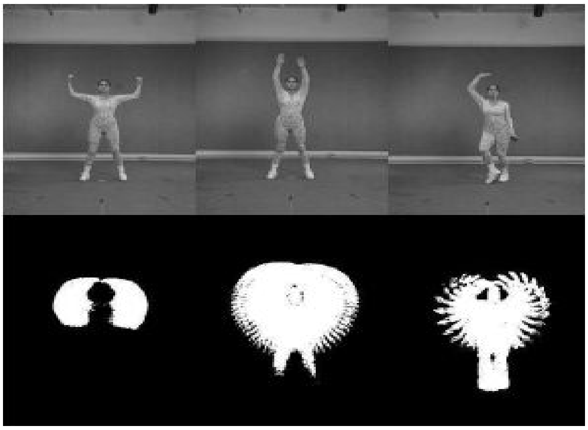


Figure 2.3 Motion energy images [18, 19]. Top row has raw data while bottom row has corresponding MEI

2.2.3.2 Motion History Images (MHI)

Just like MEI proposed in [18], Motion History Images (MHI) seek to represent a template of an action based on the tracked history of image sequences. MHI is generated from MEI. While MEI focuses on forming a template by using the spatial features to establish where the action occurred, MHI uses intensity of each pixel in the history of the motion culminating to the current activity as a function of motion density at a particular location. The MEI and the MHI can be considered a two-component version of a temporal template, a vector-valued image [4]. The main advantage that we also borrow from MHI is that the representation in MHI is done in a range of times being encoded in a single frame [24], which translates into our model descriptor. Therefore, the final MHI image records the temporal history of motion captured in a sequence of video frames. MHI is the weighted sum of the past images and the weights decay back through time. Several parameters involved in the MHI computations include position, time, and duration. Given a $MHI\ H_{\tau}(u,v,k)$, at time k and location (u,v), the MHI can be formulated as;

$$H_{\tau}(u,v,k) = \begin{cases} \tau & D(u,v,k) = 1 \\ \max\{0, H_{\tau}(u,v,k-1)-1\} & otherwise \end{cases} \qquad (2.2)$$

$D(u,v,k)$ stands for motion mask binary image from subtraction of frames and ,

τ represents maximum duration a motion is stored.

Getting MHI require some image processing techniques such as background subtraction, image differencing, and / or optical flow [22] to help identify where the motion in the current video image occurred and aid in extraction of key frames.

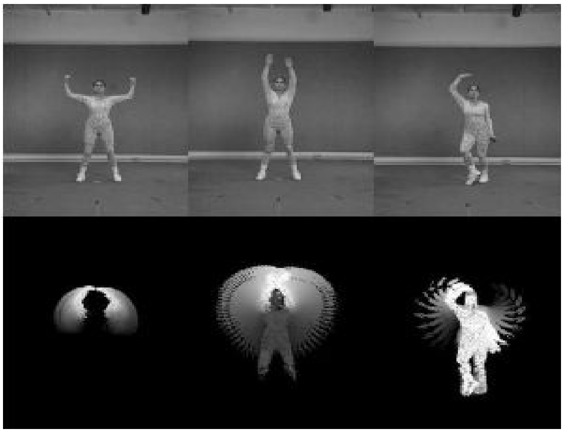


Figure 2.4 Motion history images (MHI) [18, 19]. Top row are raw data, bottom row are their corresponding MHI

Both MHI and MEI are mostly used together for better discrimination purposes [4]. MHI also help solve self-occlusion cases where direction of an action flow is vital for particular action discrimination. For instance, sitting down and standing up activities could have the same MEI representation as opposed to MHI which exhibits the direction in which the image sequence flows. MEI stores images in binary prompting some degree of ambiguity if patterns formed while the grayscale nature of the MHI shows the decay of the images with the most recent motion depicted by bright pixels. This kind of observation can tell whether the action started at a standing position or otherwise.

Chapter 3

Posture Selection and Recognition Methods

3.1 Information Entropy

This process is included in the preprocessing stages above. There is need to avoid monotony in subsequent frames from the videos. Each and every subsequent frame displays a certain change in pose of the subject in question. Information entropy aids us in determining which frames to consider as key frames whose boundary edges should be extracted for the formation of a compact descriptor. C. E. Shannon's entropy in communication theory has been widely used to detect disorder in a flow of information. A proper background on the same can be found in Shannon's 1948 A Mathematical Theory of Communication [40]. While his theory concerns point-to-point communications as in telephony addressing both source coding and channel coding, we find it very applicable in our work to help find the measure of information in each successive video frames and provide us with the disorder detected from subsequent frames. Therefore, Shannon's information entropy is used here to measure the randomness or unpredictability of a sequence of symbols. Replacing the symbols with the human action classes, the entropy value depicting difference in human body pose can be established. Shannon's information entropy can be formally stated as below.

Given a probability function, $P=\{p_{1,}p_2,...p_n\}$, defined on a discrete alphabet $X=\{x_1,x_2,...x_n\}$, entropy H(P) is defined to be

$$H(P) = -\sum_{i=1}^{m} P(x_i)\log_2 P(x_i) \tag{3.1}$$

The logarithm in the expression above normally is considered in base 2 and the units of entropy referred to as bits [41].

The entropy for a random variable X with a probability distribution P, defined on a discrete alphabet X is defined as above but with $H(X) = H(P)$. Our method computes the entropy value for each extracted window from the frame containing the extracted human silhouette pose divided into blocks of size $L * W$. Block size used is 3 * 3.

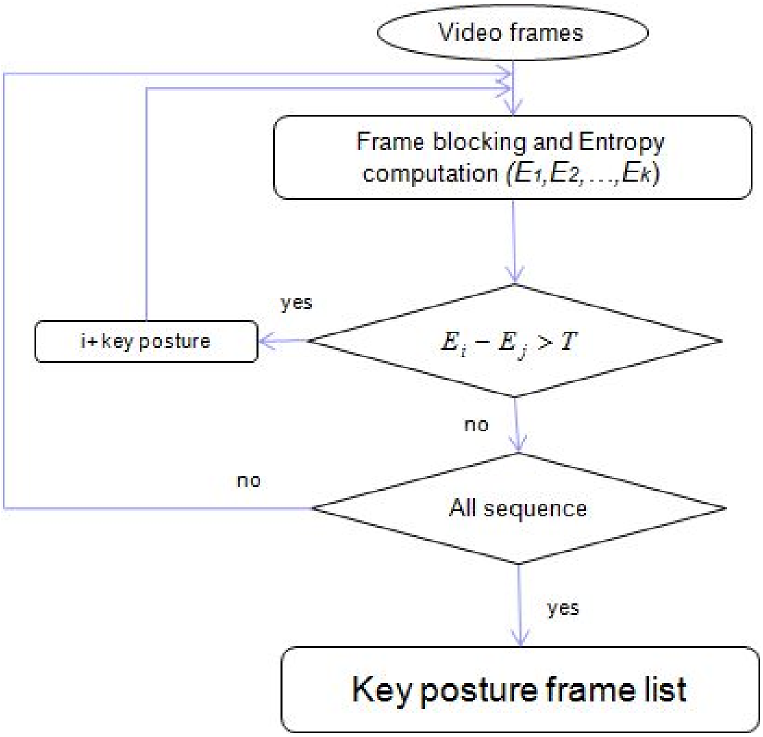


Figure 3.1 Key posture frame list formation using information entropy

3.2 Frame blocking

Entropy can be computed to determine the differences in subsequent frames. For exactness, we narrow the foreground region which mostly consist of the object for which the pose is to be determined. This is cost effective in terms of computation time and amount of data to be processed. From the preprocessing step of background subtraction above, the human silhouette is extracted on a window depending on the size of the silhouette. The window containing the subject is divided into blocks of size 3 * 3 as been mentioned above. Intuitively, the higher the number of blocks, the higher the accuracy in determining the disorder in different poses exhibited in subsequent frames. The silhouette pixels occupy the small blocks within the frames at different positions. For a frame containing a number of blocks, as in Figure 3.2, the probability of the human body pixels in each block, P_i, can be gotten by the Equation (3.2).

$$P_i = \frac{N(b_i)}{N} \quad , \ \{b_i : i = 1,2,...k\} \tag{3.2}$$

where

$N(b_i)$ represents the number of pixels of the body contained in the i^{th} block,

N is the total number of all pixels the body contains, and

k being the number of blocks the frame is divided into.

Given the equation above in (3.2), we can compute entropy of each frame from the action sequences from videos as Equation (3.3).

$$E_j(F) = -\sum_{j=1}^{m} P_i \log_2 P_i \tag{3.3}$$

F is the set of frame sequences, $F = \{f_1, f_2, \ldots f_n\}$

When two different frames of f_i and f_j are given with their respective entropies as E_i and E_j, the two frames are considered similar if their corresponding entropies are also similar. In this was, the information entropy is used to show the difference in pose or postures within video sequences. For this discrimination to happen effectively, the condition in Equation (3.4) stated below needs to be fulfilled based on some threshold comparison.

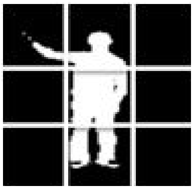


Figure 3.2 Frame blocking

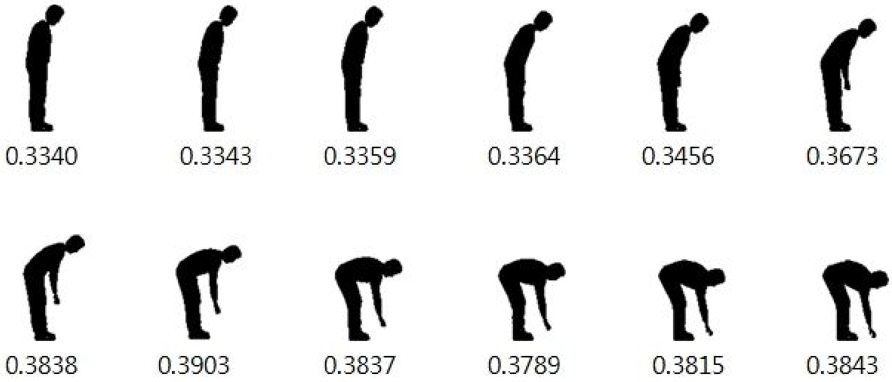


Figure 3.3 'Bend' action entropy selected images with entropy values

3.3 Shape Contexts

Shape similarities are revealed in the way objects are formed. A shape descriptor is very important in aiding for similarity correspondence computation. A descriptor is computed in an image and used to find corresponding features or points in another image. Shape context uses these features or points taken from the internal or external contours of the image. The more the number of points, the richer the descriptor will be and the representation of the shape becomes more compact and exact. The problem with this is that it increases the computation time if the full vector of a shape is used.

Shape context introduced by Belongie et. al. [42] by defining a descriptor for a boundary as a function of the other boundary points. They propose a log - polar plot of the boundary of the shape as viewed from arbitrary boundary value. Similarity between two shapes can be ascertained by finding the corresponding points and be considered to having same shape context. In one of our past work [38], we used shape context in posture matching. Example shape context is shown in Figure 3.4.

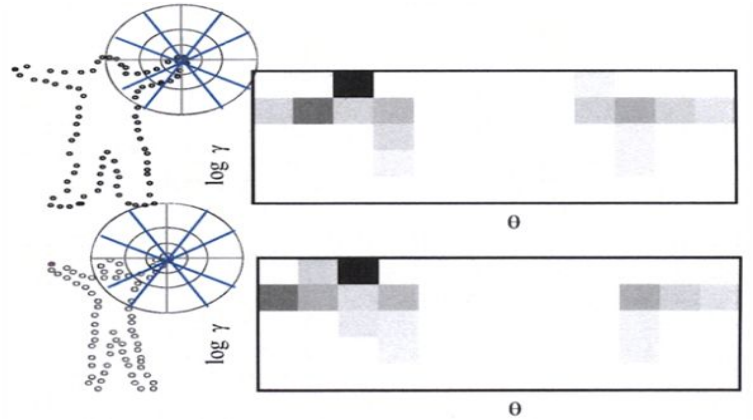
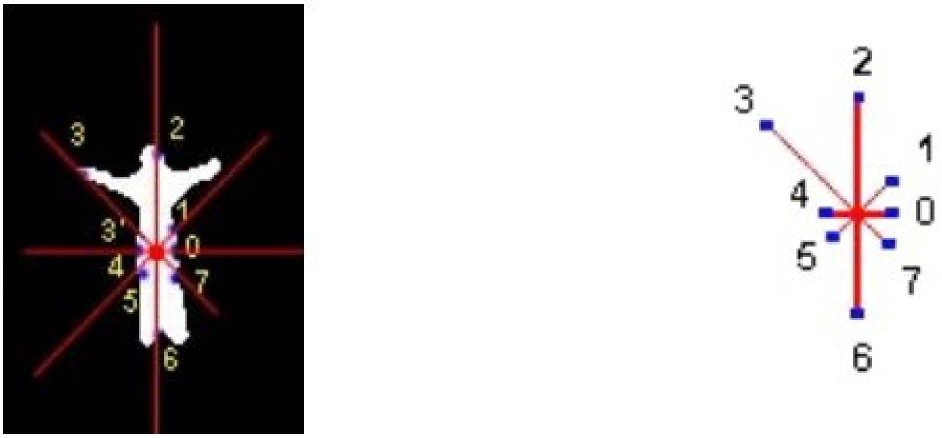


Figure 3.4 Posture matching using shape context focused on ROI

In [43], skeleton has been used to describe the shape of the human body contour. The shape description in [43] is represented by distance vector gotten from the values calculated on the skeleton of the human body contour. Here they compute the centroid of the human body image contour. Eight interested points are labeled around the centroid in eight directions of north, south, east, west, north-east, north-west, south-east, and south-west. Then the distances from the eight interested on the contour of the image to the centroid is computed to represent the image. A simple Euclidean distance is used to determine the similarity between two key postures. Figure 3.5 below shows the skeleton usage.



(a) (b)

Figure 3.5 Skeleton of a human body. (a) Interested points (b) points labelling (image adapted from [43].

3.4 Mathematical Morphology

Mathematical morphology are a group of operations that can be used in image processing in analyzing image shapes. They are very useful operations both for degraded images to help enhance the goal of operations for binary images and grayscale images. Digital images appear in many shapes and take different positions in space. Most of the morphological operations are built around reflection and translation. Reflection can be described as the mirroring on the images on a certain plane while translation is the change of position or direction detected in image as a whole or part of it (pixel level). some of the morphological operations used at some point in our framework are dilation and erosion to enhance boundary detection process. These operations use structuring elements of different sizes to be able to establish internal boundaries, external boundaries, and morphological gradients of images. Dilation, also known as Minkowski addition [44], has the effect of increasing the size of an object. Edge size can be increased or blobs in binary images could be increased to block holes in them. Erosion, or Minkowski subtraction, translates an image by reducing the image size, a process similar to thinning. This process erodes an image and if applied continuously could result in complete disappearance of the region of interest. Opening and closing are other morphological operations that are useful in noise removal. Hit-or-miss transform and skeletonization operations are also worth mentioning for their contribution in image processing [44].

[45] explains the fundamental advantages of morphological operation as intuitive and works directly on the spatial domain of the image. In looking for an appropriate descriptor to analyze our action patterns, this task calls for the image transformation into other domains which can be read by machines

more easily and recognition procedure be carried out efficiently. We have transformed the morphologically operated model images into polar logarithmic representation for mapping them onto each other for recognition and classification. We give a brief overview of log-polar transformation below.

3.4.1 Sobel edge detector

As suggested by the title of this work, edge is very vital part of the feature to be extracted for our descriptor. We define as edge as abrupt change in color intensity on image surface. Sobel edge detection [58, 59] uses convolution kernel to create a series of gradient magnitudes in horizontal and vertical directions. The goal of the Sobel operator is to pass through the image and find where the intensity varies from low to high and vice versa. To detect the edges, the operator uses two convolution kernels to detect both the vertical contrast (h_x) and horizontal contrast (h_y).

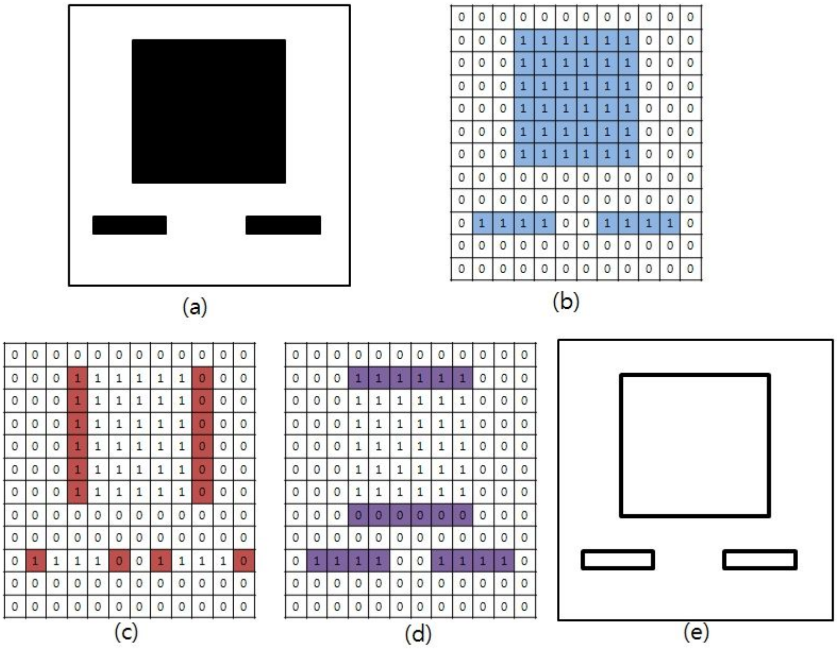
$$h_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad \text{and} \quad h_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

To be able to create a series of gradient magnitudes, the images $x[,]$ and $h[,]$ are convolved to give a product image $y[,]$. The size of $h[,]$ is $M \times M$ pixels with the indexes running from 0 to $M-1$.

$$y[r,c] = \sum_{k=0}^{M-1} \sum_{j=0}^{M-1} h[k,j]\, x[r-k, c-j] \tag{3.6}$$

In Equation (3.6), individual pixel in the output image $y[r,c]$, is calculated according to the side view. The indexes j and k are used to loop through the rows(r) and columns(c) of $h[,]$ to calculate the sum-of-products.

We give a simple sketch of what happens to a binary image when the two convolution kernels above are passed through. The image is generated for the sake of understanding.



(a)

(b)

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(c)

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(d)

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	1	1	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(e)

Figure 3.6 Illustration of Sobel edge operator.

Figure 3.6 (a) is the original image whose edge is to be detected, while (b) is matrix representation of (a). (c) and (d) are the outcome after horizontal and vertical scans by h_x and h_y, respectively. (e) shows the final edge map as extracted by the Sobel operator.

3.5 Log-polar transformation

This is a simple transform operation that changes the coordinate system of a given image from cartesian to log-polar. Our image descriptor is in log-polar form for comparing both the probe and the query images during classification. We view this step in our framework as very important for image representation ready for similarity measurement. In pattern recognition, two images to be compared should have similar properties like size in order for comparison to be on equal level ground or share some other intrinsic features that can be observed. The two images to be compared should have gone through proper registration process and be fairly close in scale, rotation, and translation. The task to recognize actions require keen registration procedure due to the dynamic actions performed by different people of different sizes and heights; young and old, big and small, tall and short. These external influences introduce variations that could hinder classification process and produce higher false recognition rate thereby reducing the performance of machines.

Log-polar registration [46, 47, 48] will bring images into close alignment even where large scale changes, arbitrary rotations, and translations are experienced. This property is very important for us in this work to convert an image to a form that is rotation and scale invariant. A two dimensional mapping is shown in Figure 3.7 with the following Equations in (3.7) and (3.8) to compute the rho(r) and the theta(θ).

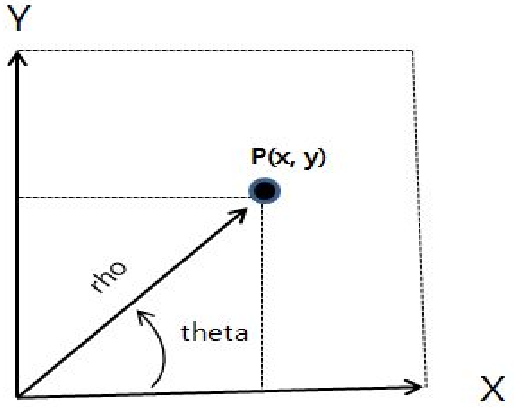


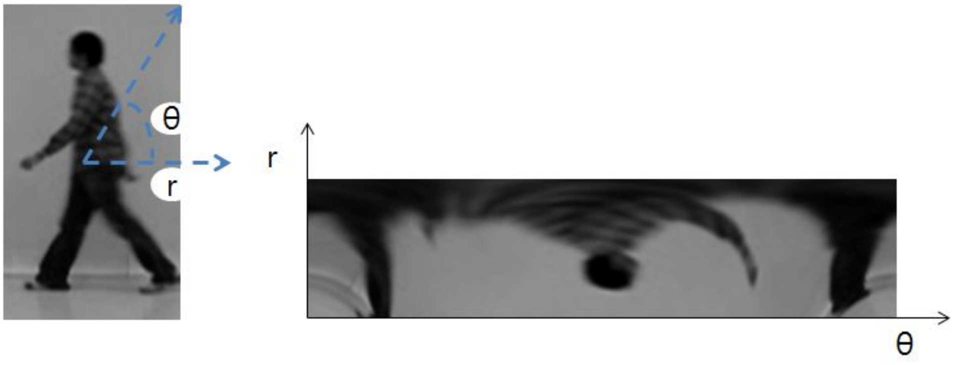
Figure 3.7 Two-dimensional mapping into polar coordinates

To describe the log-polar coordinates operation [49, 72], lets consider the polar (r, θ) coordinate system where r is the radial distance from the center (x_c, y_c) and θ to denote angle. Any (x, y) point can be represented in polar coordinates as;

$$r = \sqrt{(x - x_c)^2 + (y - y_c)^2} \tag{3.7}$$

$$\theta = \tan^{-1}\left(\frac{y - y_c}{x - x_c}\right) \tag{3.8}$$

If polar coordinate transformation is applied to an image I, it maps radial lines in cartesian space to horizontal lines in polar coordinate space. Pictorial view is given in Figure 3.8. This transformation enables multi-resolution feature analysis in both radial and angular directions which best describes the action shapes clearer and better for discrimination purposes.



(a) (b)

Figure 3.8 (a) The original input image, and (b) is the log - polar transformation of (a).

It is worth noting the importance of the two parameters r and θ in the Equations (3.7) and (3.8) above. The parameters can be adjusted depending on the image being operated on and the purpose of doing the transformation. Radial axis, r, determines the field of view of the transform by giving the radius of the region of the image to be sampled. The angle, θ, determines the resolution as seen on the outside ring. With the right adjustment to these parameters, the full detail of the original image can be captured. Empirically, the value θ, was kept smaller to avoid oversampling of image pixels at the centre of the image.

This was implemented by MATLAB Toolbox where there is a function to manipulate the images to be operated on. The function transforms an image from cartesian to polar with the function, *cart2pol(x, y)*, which transforms two-dimensional Cartesian coordinates stored in corresponding elements of the given arrays *x* and *y* into polar coordinates. MATLAB Toolbox further compute the log-polar transform of the image by manipulating the images through the $[img, rMin, rMax, M, N]$ parameters. These represent the image, minimum radius, maximum radius, and M and N being the number of pixels in the rectangular domain, respectively. Of importance is the establishment of the center of the pattern in the original image we intend to do the operation on.

3.6 Action Recognition Methods

The previous sections were dedicated to extracting and modeling the descriptor for the sole purpose of this step; recognition and classification into preferred categories. Recognizing patterns is a very sensitive part of computer vision and in our framework too because it shows whether our work is completed successfully or not. Many methods have been put forward depending on the tasks they are designed to accomplish. Complexities of the images to be processed influenced by the background and multi-actions preformed in a crowd is a big challenge even to our own method and is the direction the future research will take. One basic task they all have sought to achieve to this moment is to determine the extent to which a shape (query) differs from another(probe). Methods like Support Vector Machines (SVM) [50, 51] have been applied successfully in some cases. Correlation and template matching methods [52] are also other examples of recognition which have been applied. Other methods involve model-based vision techniques and have been explored.

We have settled on Hausdorff distance method due to the nature of our final descriptor which is in a log polar transform for the reasons that we explain in section 3.6.1. Another intriguing method that we carried some experimental test with is the Chamfer matching method.

3.6.1 Hausdorff Distance Measure

Hausdorff distance (HD) measure [53] is a distance measure that computes the maximum distance of a set to the nearest point in the other set. In other words, it measures the extent to which each point of a model pattern set lies near some point of the probe image set. With this value computed, we can use it to determine the similarity between the two objects / images in question. This is done by superimposing the images on one another. This is the characteristic that we need for the comparison between the action trained descriptor models and the input image for classification.

Hausdorff distance method has been widely used and is lauded for many advantages with the two main ones enhancing the effectiveness of our proposed descriptor.

i). HD is simple to implement and has a fast computation time. The log polar transforms of our descriptor are of the same size and superimposing makes comparison easy even visually.

ii). Our work involves stacking of frames to form a summed up edges for description. HD is relatively insensitive to small perturbations of the image such as those that occur with edge detectors and other feature extraction methods.

Hausdorff distance can be extended to carter for occlusion and help identify portions of a shape hidden from view but need to be identified. Occluded images are not included in our work but HD could help solve that in case we are faced with it.

We formulate the HD method of recognition as in Equation (3.9). Given two finite point sets of $A = \{a_1, a_2, ..., a_p\}$ and $B = \{b_1, b_2, ..., b_q\}$, the Hausdorff distance is defined as;

$$H(A,B) = \max(h(A,B),\ h(B,A)) \tag{3.9}$$

where

$$h(A,B) = \max_{a \in A} min_{b \in B} \parallel a - b \parallel$$

and $\parallel . \parallel$ is some underlying distance on the points of A and B, in this case, we assume it to be Euclidean norm.

HD is asymmetric, which means that the condition $h(A,B) \neq h(B,A)$ could apply. In comparing two images, each vertex on both images must be compared with all the others on the other image and their distance value computed. The minimum or maximum values are not the same in some cases, and that is the reason for the condition above of inequality. The distance $h(a,b)$ is actually the largest of the two minimum distances of $h(A,B)$ and $h(B,A)$. This argument put above can be written as in Equation 3.10.

$$h(A,B) = \max_{a \in A} \{\min_{b \in B} \{d(a,b)\}\} \tag{3.10}$$

where a and b are points of sets A and B, respectively. The $d(a,b)$ is taken to be the Euclidean distance between a and b, for simplicity reasons.

By computing the maximum distance between two sets of images, Hausdorff distance measures the mismatch between the two sets which are at fixed positions with respect to each other. This is true as our images that we intend to compare are fixed without translation. If translations were to be considered, HD obeys metric properties as can be seen in [54, 55]. This means that the function is everywhere positive and has the properties of identity, symmetry, and triangle inequality. We hold the view that a pattern formed from modeling a particular action could only resemble the same pattern executed by the same action. On the same note, the order of comparison of the formed patterns does not matter and that two different patterns can not all be similar to a third one. All these notions are inline with the three properties of identity, symmetry and triangle inequality mentioned above.

3.6.2 Chamfer Matching Distance Measure

Chamfer matching [56], is a matching technique that compares the shapes of two collections of shape fragments, at a cost proportional to linear dimension, rather than area. N. G. Barrow et. al in [56] stated that chamfer matching exploits more knowledge of the invariant 3D structure of the world and the imaging process. This property can be very useful for our descriptor which relies so much on spatio-temporal aspect of the action performed and the features extracted that form the shapes. To compare or to find similarities between objects, several methods either use low-level information such as pixel value, or high-level information like edges. Hierarchical Chamfer matching [57], an improved chamfer method is relatively insensitive to noise, rotation, scaling and other small perturbations experienced. We view this work to be practical in our work which involves comparison of individuals of various sizes carrying out similar actions.

Chamfer distance can be used to measure similarities between two images. Given the two point sets $U=\{u_i\}_{i=1}^{n}$ and $V=\{v_j\}_{j=1}^{m}$, the chamfer distance function is the mean distances between each point, $u_i \in U$ and its closest point in V. This is formulated as in Equation (3.11);

$$d_{cham,\tau}(U,V)=\frac{1}{n}\sum_{u_i\in U}\max\left(\min_{v_j\in V}\| u_i - v_j \|,\tau\right) \tag{3.11}$$

where τ is the threshold value, reducing the effect of outliers and missing edges.

Usually the chamfer distance function between two shapes is efficiently calculated using a distance transform. In our experimental results, distance transform of our edge model has been computed and image gradient gotten before their log-polar transformations.

Distance transform (DT) takes a binary feature image as an input and assigns to each pixel in the image the distance to its nearest feature.

3.7 Distance transform

This operation on the binary image pixels assigns the distance value between a pixel and the nearest nonzero pixel. The distance can be achieved through methods like cityblock, chessboard, euclidean, and / or quasi-euclidean. Euclidean distance is the most commonly used method and is described between two pixels (x_1, y_1) and (x_2, y_2) as Equation (3.12).

$$d_{trans} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.12}$$

(x, y) being two points on a 2D image plane.

To enhance the images and bring out additional hidden details for the transformed images, *image gradient* [59] could be computed. Image gradient describes the continuous shift from light to dark, or dark to light spectrum of the image. These images could be RGB or binary. Our images belong to the latter. The shift is all about the intensity of a single color, especially for our case which is binary. The color intensity goes from white which is the maximum intensity, to black, depicting minimum intensity. The image gradients are useful in determining the vector direction and also the speed of color shifting objects. They are also used as a description of the shift in intensity and amount of light reflected by a color within a single color. As stated above, high intensity results from a color reflected towards the viewer and is closer to white. Low intensity means that the color absorbs more of the light and is therefore closer to black in color. This operation applied on the image surfaces covered in tiny gradients is a smoother version of the same. This characteristic of showing different intensities depending on reflection or absorption value makes the pattern readable for the computer to recognize between different gradient areas adjacent to one another. This is a very important information needed for edge discrimination and thereby help derive a digital representation of an image. The formulation in computing the image gradients have some similarities with the edge detection methods in principle. The gradient of an image is given by the Equation (3.13);

$$\nabla f = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} \tag{3.13}$$

where $\frac{\partial f}{\partial x}$ is the gradient in the x direction,

while $\frac{\partial f}{\partial y}$ represents gradient in the y direction

The overall gradient direction is calculated by the Equation (3.14).

$$\Theta = atan2\left(\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right) \tag{3.14}$$

Chapter 4

The Silhouette edge-based Log-polar Descriptor

4.1 Introduction

As been stated in the previous section, we present a silhouette boundary edge-based approach for locating, describing, and recognizing human actions in videos. Locating subjects to be recognized in videos is a task that researchers have done a great job on. The subjects that have been located and extracted in video sequences need to be described in a way that they are uniquely represented and should be distinguishable from other activities. This is not a so trivial task to undertake in computer vision. There are many adversaries that make it difficult to achieve this goal that we will discuss later in this chapter. A properly described subject is stored either as a template or an exemplar to be used at the categorization stage for different actions. The similarity measurement is used to group actions into their appropriate categories.

We seek to derive a robust descriptor with very salient features that adequately and correctly represent the various target actions as seen in the videos. In the end, the goal is to be able to label image sequences in a video with action labels that reflect the activities happening in videos. A sequence labeled as running should contain the running activity, and the other various activities should also be the same. The training of the template to be used as models in the gallery are trained offline. The verification process could be applied both online and offline. In the end the actions in a video will be localized in terms of extracted featured which are spatio-temporal in nature. The extracted features can finally be used to recognize the activities in a video sequence.

4.2 Framework of the proposed Method

The proposed framework consists of four main parts; preprocessing, feature extraction, representation, and recognition. We look at each of the processes shown in Figure 5.1 in the context of their contribution to this work. The main parts of the process are geared towards the development of a compact representation of activities in videos through a log-polar transformed template.

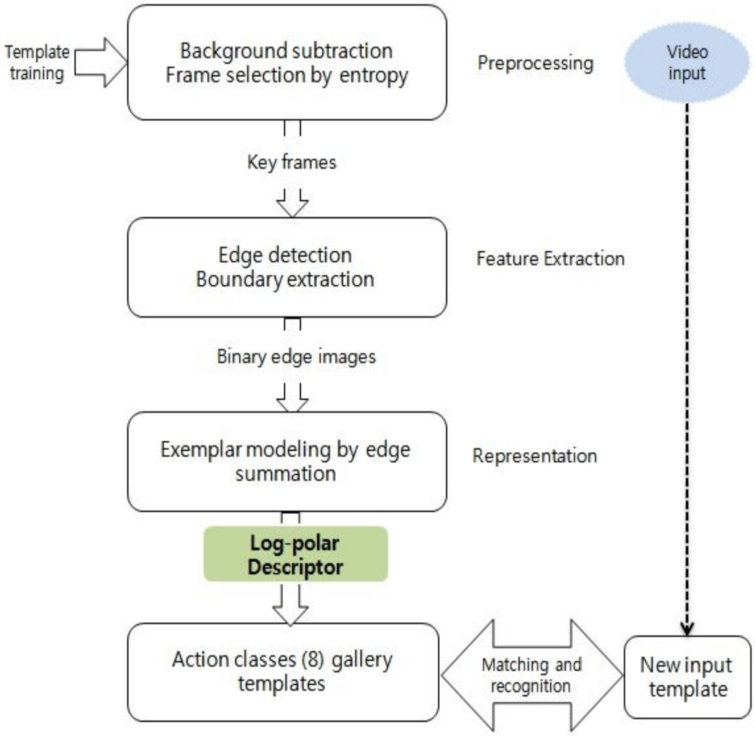


Figure 4.1 Framework flow.

4.2.1 Preprocessing

Proper image representation requires some kind of preprocessing in order to extract the right features for computer use. This is a necessary step that involves selection of key frames that we plan to use in forming the descriptor both for training and input data. In action recognition, the movements in the videos is very vital to be clearly noticed. The backbone of the research here is dependent on the orientation of the body parts that would depict a kind of activity. Therefore, the movement of the body is very important for us and all efforts should be invested in extracting every bit of this information. The movements we focus on usually take place at the primitive level; that is, the human limbs do move a lot more that the human torso or even the head. There is need to capture any slight movement within these three levels as they occur in subsequent frames in videos. The difference in the pose or posture due to limbs movements can be captured and used to show that an action has occurred different from the previous one.

One very common and imminent challenge that we need to address here is the effect of background during human feature extraction in videos. The task become even more tedious in cluttered scenes and in multi-activities by crowds. This enormous task has been significantly reduced in our research work due to the constrained public databases used because of the uniform background. The database we produced ourselves was also done in a very controlled environment. The backgrounds are static in nature with constant illumination effect and typical human actions being performed. This challenge aside, we still have to separate the subject from the background for salient feature extraction. This is a process of separating out foreground objects from background in a sequence of video frames. Background subtraction [25, 26] with repeated binarization is used to accomplish this step in our work. A part from background subtraction, other methods considered for motion segmentation but not used in this work include frame differencing [27, 28], adaptive median filtering [29], median filtering [29], mixture of Gaussians [29], and Kalman filtering [30]. Figure 5.2 (a) and (b) show the original data and some of the results of preprocessing before the feature extraction process, respectively. The results shown were empirically conducted on two of the three databases used in this work.

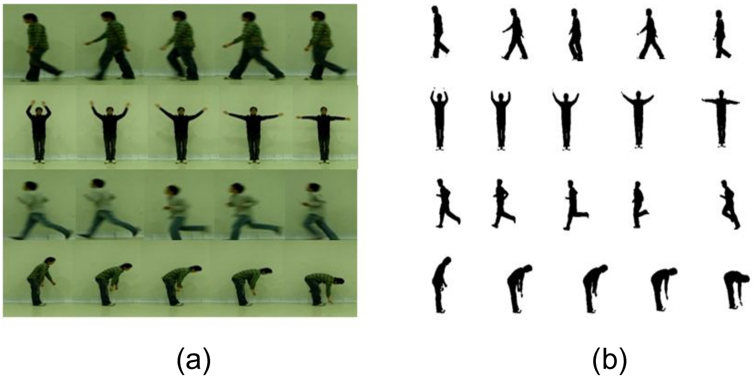
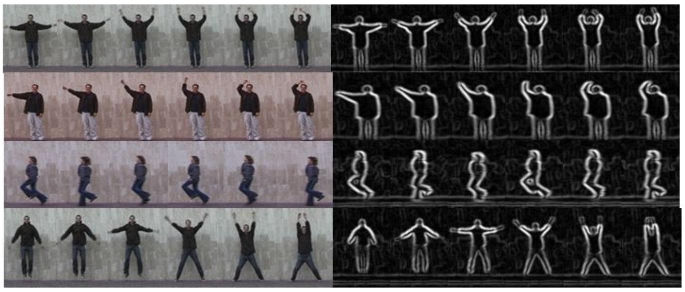


Figure 4.2 (a) original video sequence images from own-generated database (OGD). (b) selected images with background subtracted in the first step of preprocessing.

In Figure 4.2, part (a) is the raw data taken from our own-generated database. Part (b) shows example binary images after background subtraction process applied. Figure 4.3 shows Sobel edge detection [58, 59] as another way of separating the foreground objects from the background scene. The images experimented on are from Weizmann Human Action database [37]. Subsection 4.2.2 on feature extraction reveal that the edges extracted from binary images after background subtraction perform just as the edges by Sobel edge detection. In the case where Sobel edge detection is used, the two steps of preprocessing and feature extraction can be combined into one. This is because Sobel edge detection can be applied as a feature extraction technique, too.

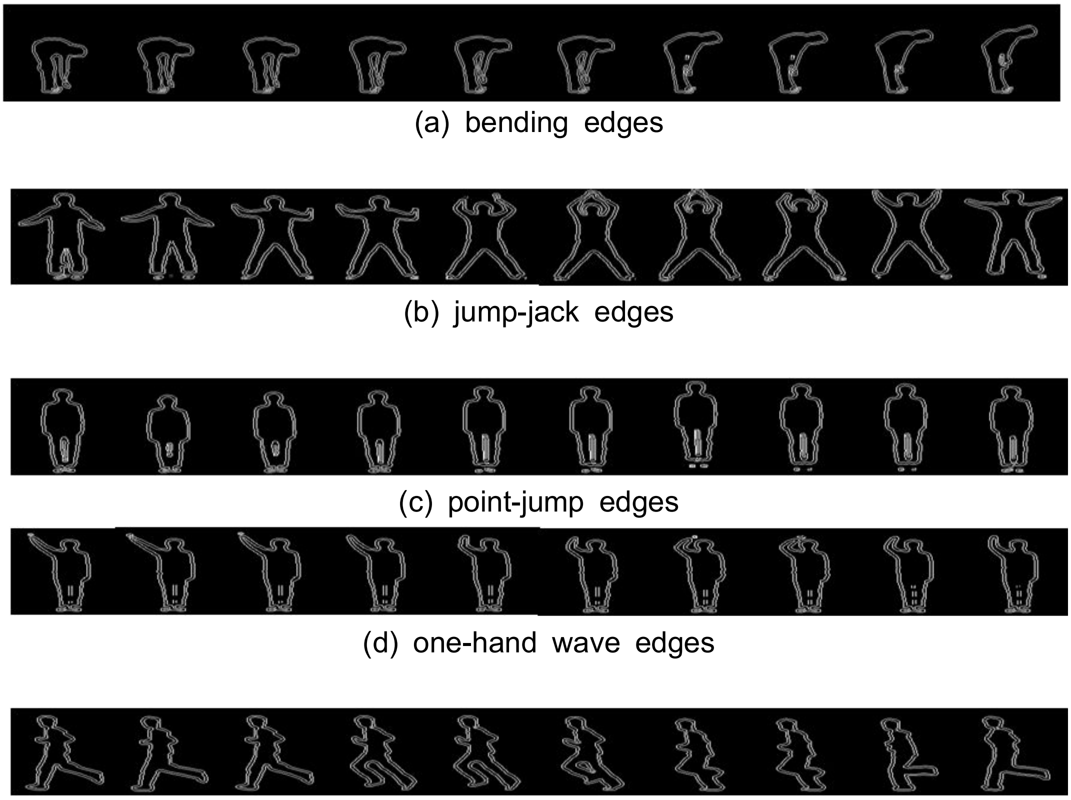


(a) (b)

Figure 4.3 (a) Raw data selected from Weizmann Database, (b) edge detected images through Sobel edge detection mechanism

4.2.2 Feature Extraction

A very important step that defines whether the action is going to be properly represented or not is the feature extraction process. Simply put, preprocessing step above enhances the extraction process. The feature we chose depends on the global boundary edges of the human silhouette separated from the background in the preprocessing step. Edge detection, as expounded on in chapter 3 subsection 3.1.4 of this work, is a significant process that aids in locating the edges from the extracted foreground images. The edge orientation is very important for our descriptor for the understanding of the image features and they contain significant information. Some other methods may chose to use features like color or texture. Using the edges has enabled to reduce the amount of data to be processed by a wide margin while still maintaining the vital information needed for data representation. The subsection on the image gradient in chapter three support this fact. The high frequencies exhibited on the edges is an asset here [39]. The structural properties needed for successive processing have been preserved while image size has been reduced significantly.



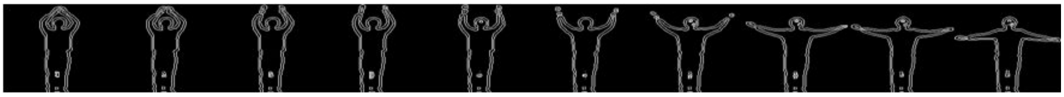
(a) bending edges

(b) jump-jack edges

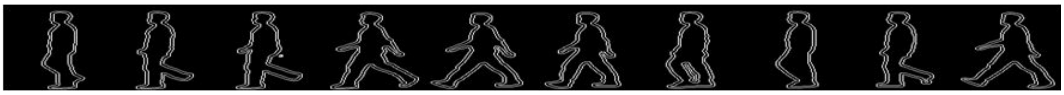
(c) point-jump edges

(d) one-hand wave edges

(e) running edges



(f) two-hands wave edges



(g) walking edges

Figure 4.4 Sequences of boundary edges belonging to various actions as indicated from (a) to (g).

The boundary edges extracted from human silhouette from Own-generated database, has been shown in Figure 4.4. The same can be seen by the example edge images from Weizmann human action database in figure 4.3(b) on preprocessing.

4.2.3 Representation

The proposed descriptor is a form of representation from the extracted edge-features making this subsection part of the central contribution and focus in our work. The extracted features from the video frames are represented and described in a suitable form ready for further computer processing. A descriptor is the yardstick used during the process of discrimination of the actions either as similar or dissimilar in characteristics displayed. The recognition process will need a standard measure, a descriptor, to compare the probe images with. Recognizing different actions is more of solving a shape problem. This challenge is mostly solved based on the external characteristics rather than the internal characteristics of the object to be recognised. External here refers to the shape boundaries /

edges extracted from the human silhouette. Internal involves the consideration of pixels found in the regions covered by the object. The silhouette boundary-edge that we propose encapsulates both the external and external properties of the activity from the video sequence. The boundary edges represent the shape extracted while the internal property is represented by the temporal energy due to body or limb movements. The temporal energy is very dominant on areas with large movements and is reflected on the model images by the high pixel concentration in particular regions. The superimposed boundary edges extracted from the key posture frame list form a unique shape / pattern that is used to describe different actions that we intend to model and recognize. The combination of the two properties of representation constitute part of the main contribution of this work.

Nazli et. al. [60], propose the usage of histogram of oriented rectangles as a pose descriptor for human action recognition. Their feature extraction commence in a similar way like ours by first subtracting the background to separate the human figure in the frame. They then search for rectangular patches on the extracted silhouette on the whole body, limbs included. Histograms of oriented rectangles are then computed on different regions from the equally separated grids on the silhouette by the bounding box. Their descriptor is based on the combination of these histograms from each subregion. Figure 4.5 illustrates how histogram of oriented rectangles is computed from [60]. Here the bounding bos around the human silhouette is divided into an N * N grid. The histogram of oriented rectangles from each spatial bin can be seen in the figure too.

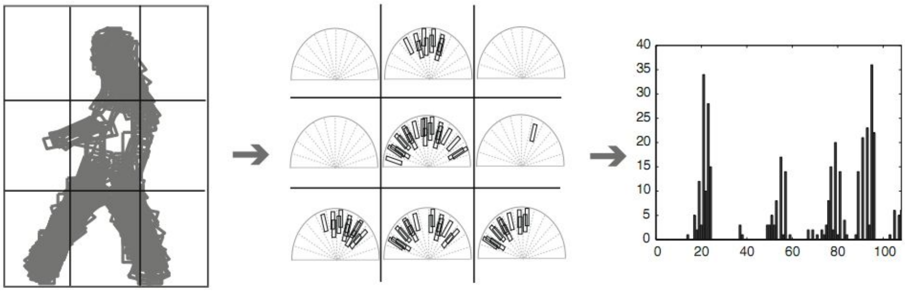
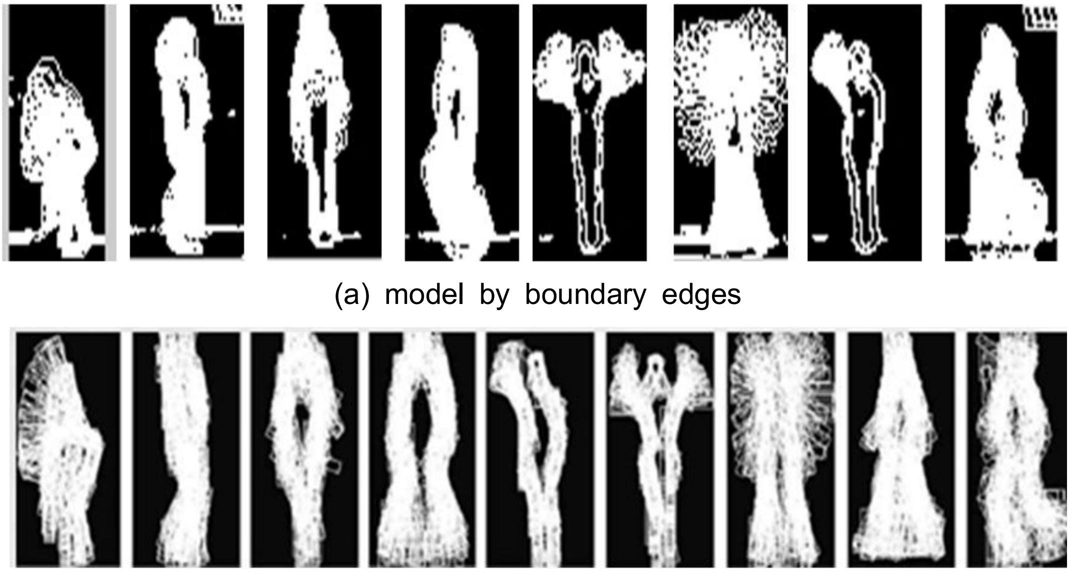


Figure 4.5 Histogram of oriented rectangles [60].

Of our interest is the summation of image sequences leading to the formation of the spatial histograms of oriented rectangles globally. The superimposed boundary edges we use reveal much similarity with this method in the globally produced patterns for different actions. The same pattern is seen to form from across the three databases used in our experimental analysis. The visual comparison of the global model suggest that a robust feature descriptor can be formed to expediate the discrimination process. In Figure 5.6, we show this visual similarity existing between the block system used by Bobick et. al. and our boundary-edge model.

The Figures in 4.6 do not compare one on one the images in them but it is clear that though the methods of extraction was different in order to reach the modeled actions, similarities can be easily spotted between actions of the same grouping. In Figure 4.6(a) the actions are from left to right 'bend', 'front-jump', 'point-jump', 'forward-skip', 'two-hands wave', 'jack', 'one-hand wave', and 'run', respectively. Figure 4.6(b) from left to right are bend, jump, jump-in-place, gallop sideways, one-hand wave, two-hands wave, jump jack, walk and run actions.



(a) model by boundary edges

(b) model by blocking [60]

Figure 4.6 Some similarities revealed in shape pattern by boundary-edge summation method and histogram of oriented rectangles. Both methods stem from MEI [18]

Where as [60] chose to used histograms to describe the patterns, we opted to implement more image processing to come up with better and simple patterns that could be compared easily against each other. Distance transformation and log-polar transform has been explained above to form a unique descriptor for human action representation.

We give an example representation of a two-hands wave actions developed from our own-generated database in Figure 4.7. The process of recognition is explained in detail in subsection 4.2.4 where the final descriptor is used as a template in the database to represent many other actions modeled. Figure 4.7 flows from the superimposed edge-images, then to distance transform of the image, and finally the log-polar transform of the model.

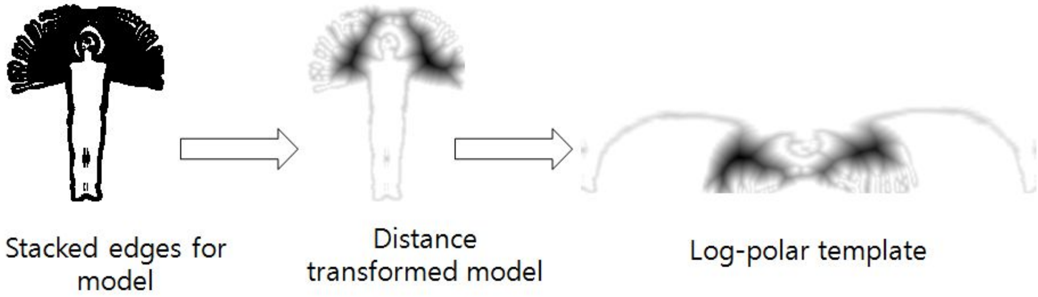


Figure 4.7 Two-hands wave representation as log-polar template.

4.2.4 Recognition

The three subsections discussed above are a build up to this last part of recognition. The sole purpose of forming the patterns from above procedures is to categorize them into different groups. In our work, we have described and trained eight different kinds of recognizable activities ready for the matching with the new input images. The method can be used to describe and represent many more others activities. Matching in this case deals with transforming a pattern / image and measuring the resemblance with another incoming pattern / image using some dissimilarity measure. We explore Hausdorff distance measure to reveal any similarity that may exist between two images. The result is compared with the performance of other distance measures like Chamfer matching [56], Support vector machine (SVM) [61, 62, 63, 64], Mahalanobis distance [65, 66], K-nearest neighbor [67, 68, 69], Multi-class nearest neighbor [61, 64], and maximum likelihood [70]. Squared Euclidean and correlations as measures of similarities have also been applied by some researches. They all seek to classify the unrecognized motions into their proper classes / grouping. The full transformation of the silhouette-edge model to recognition process according to our method would look like what is shown in Figure 4.8. We have taken the act of 'jump - jack' from it's edge model to the final log-transform shape to be compared with a log transform of a 'bend' action.

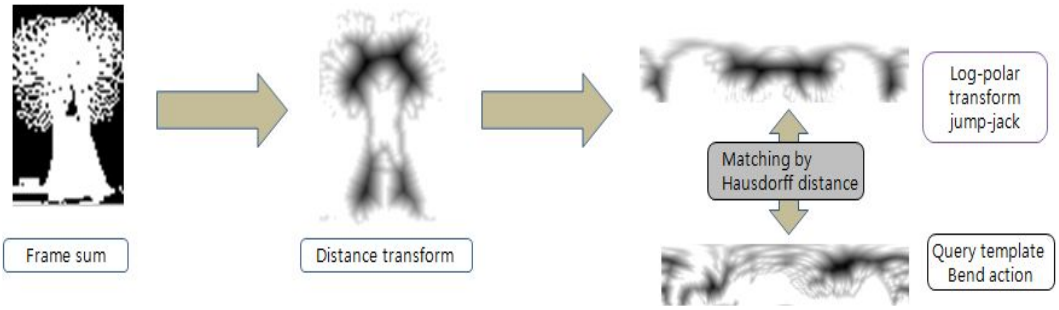

Frame sum
Distance transform
Log-polar transform jump-jack
Matching by Hausdorff distance
Query template Bend action

Figure 4.8 Two transformed images being compared with each other.

4.3 The proposed feature descriptor step-by-step.

As we seek to create a robust descriptor for our action recognition process, we cannot help but think of this process as similar to that of extracting biometric traits for identification purposes. A unique descriptor should be no different from the qualities of universality, distinctiveness, permanence, and collectability that biometric techniques require. Every action to be described should be universal among all subjects. That is, a 'run' is a 'run' whether carried out by a small person or a big person. This quality applies to all other actions that you would wish to train and classify. Distinctiveness in that no two actions should have the same features is vital. This quality is very challenging especially in self occlusion situation. The final descriptor should be able to overcome that. Permanence is another quality that ensures the information extracted is very stable for a long period of time. The last point is very pivotal for quantitative analysis of the action in question. It asks the question whether the movements observed are collectable and if they can me measured for comparison purposes.

As been explained in the previous chapters, the task we are trying to accomplish involves a combination of procedures that finally boils to the proposed descriptor. In the process flow of the framework from Figure 4.1, image preprocessing, feature extraction, representation, and recognition were combined to complete the task. We lay out the flow of the framework again as follows;

(i) Given the input video, motion analysis is performed in the video to extract the object, in our case the human body, in each of the frame sequences. Background removal is done here simultaneously for the foreground (human silhouette) to remain.

(ii) Information entropy from chapter 3 is applied for key frame list composition.

(iii) From the selected key frame, edge detection is performed and extraction of both internal and external boundaries executed.

(iv) The finely selected boundary edged are superimposed to form a pattern that each action can identify with. This procedure done with all the classes and a unique shape formed per class.

(v) The shapes formed from (iv) undergo through distance transformation for clarity and readability improvement before getting their log-polar transform. The merits of log-polar has been explored in section 3.5 above.

(vi) Action classification can now take place through Hausdorff distance measure.

The steps laid above lead to the formation of the algorithm below. This pseudo-algorithm contain different parts as has been explained deeply within this work. The end result is a log-polar descriptors that can be used to classify various actions modeled as per the user. Some explanation of the notations in the pseudo-algorithm;

img_edge stands for extracted edge images from the key frame list.

i represents the different types of actions modeled.

n is the total number of actions and can also be used to denote the total number of modeled edge images or templates.

dist_imgs is short for distance images; images that have undergone distance transformation.

HD is the Hausdorff distance measure

templates are the modeled log-polar images stored in the database while *input* is the input template to be categorized.

log_polar is the function that transforms the images into polar images

edges is a function that extract edges from silhouette boundaries

dist_transform is a function that computes the distance transform of the images.

Algorithm: Log-polar template matching

```
1. Extract the edges from the entropy computed video sequences in the key list
        $img\_edge = edges(key\_frames)$
2. Build a model by edge summation
        $for$ actions, $i = 1, 2, \ldots, n \Longrightarrow \sum_{i=1}^{n} i$
    $n$ = total number of actions equals to number of model templates
3. Get the distance transform of $n$
        $dist\_imgs = dist\_transform(i)$
            $i++;$
4. Transform distance images into log−polar form
        $templates = log\_polar(dist\_imgs)$
5. Compare the two log−polar forms of templates and the input images through Hausdorff distance measure
        $HD(templates, input)$
6. Assign the input to the template with the minimum distance
        $action = minHD( template, input)$
        where $template = 1, 2, \ldots, n$
```

Figure 4.9 The proposed feature descriptor Algorithm

Chapter 5

Experimental Results and Analysis

5.1. Introduction to databases

Publicly available databases are mostly used to validate the many different approaches and provide the platform for performance evaluation and comparison. Databases also come with various test environment and provide researchers with a variety of challenging scenes and complex actions to be recognized. Some databases contain indoor activities, others outdoor activities, and while others depict complex scenes with multi-activities going on. All these are very vital for specific research problems. This dissertation mention quite a number of databases but concentrate on the three that are mostly used for our method's validation.

Hollywood human dataset [32, 33], Figure 5.1 (a), has 8 and 12 different actions extracted from movies performed by various actors. You realize it is quite challenging just by the fact that actions are extracted from movie scenes which are bound to face hurdles like dynamic backgrounds, occlusions, and movements of camera.

UCF sports action dataset [34] shown in Figure 5.1 (b), contains about 150 sequences of motions in sports including actions like kicking, swinging in baseball, diving and skating among others. Here too variations in action performance, human appearance, viewpoint and illumination are factors that pose great challenge. INRIA XMAS Multi-view dataset [35] is among the many databases with interesting images and equal challenges to work with.

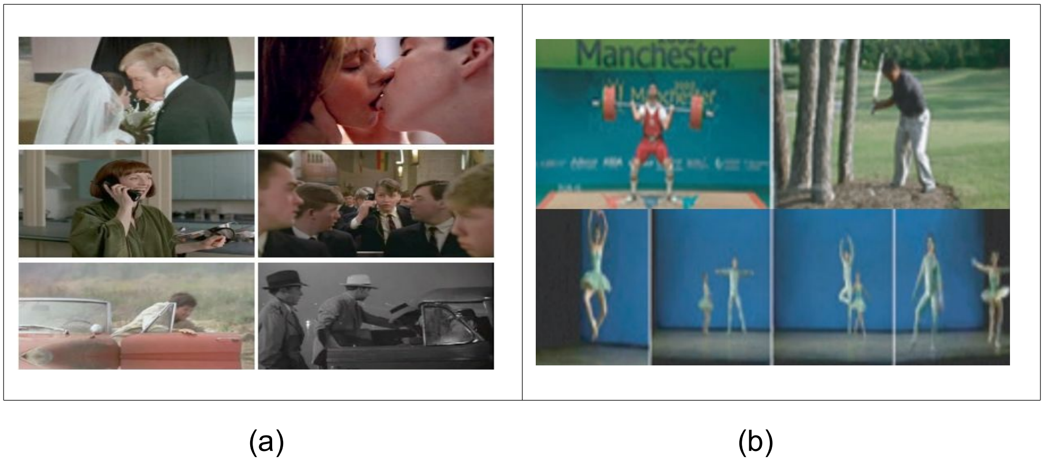

(a) (b)

Figure 5.1 (a) Hollywood human action dataset while 5.1 (b) are some captions of sports activities from UCF sports action database reprinted from the references above

We have used two main publicly available databases and one own-generated dataset. The explanation of these three databases follow down here.

5.1.1 KTH human motion database

This dataset formally introduced by Schuldt et. al. [36] consists of six challenging classes of human actions including boxing, hand-waving, running, jogging, walking, and hand-clapping. This dataset contains one action per video and a total of 600 video sequences in it. 25 subjects perform the actions in four different scenarios; outdoors, outdoors with scale variation, outdoors with different clothes on, and indoors. A snap shot of KTH dataset is shown in Figure 5.2. Visually, it reveals the close variations between the act of jogging and the act of running that confuse a lot of machines, including human eyes. In this dataset, we used 15 subjects for training and 10 subjects for testing.

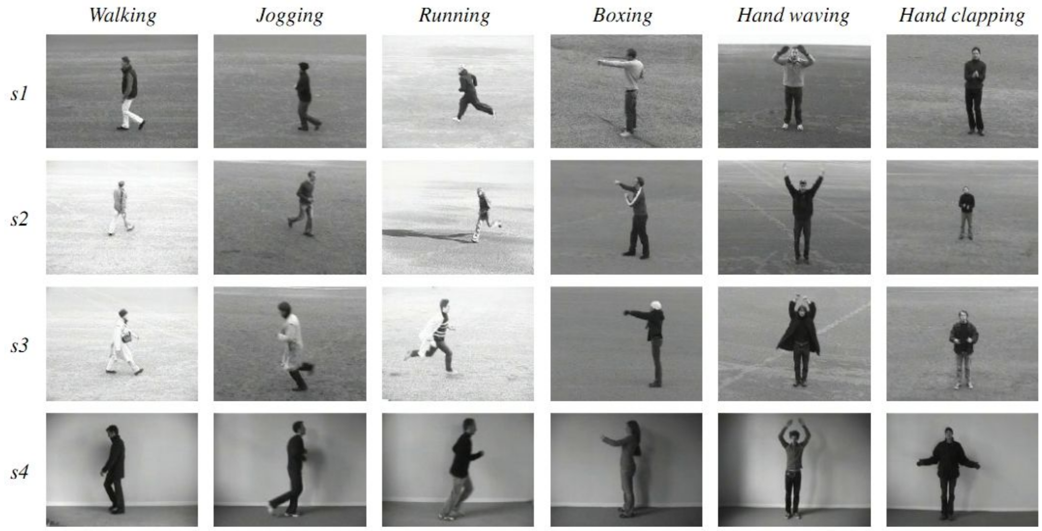


Figure 5.2 KTH database in 4 different scenarios. *Rows:* Outdoors (s1), outdoor with scale variation (s2), outdoor with different clothes (s3), and indoors (s4). *Columns:* Different actions as indicated.

5.1.2 Weizmann database

Weizmann Institute of Science's human action database [37] is introduced by Blank et. al. The database contains a set of ten actions including bending, galloping sideways, jumping, jumping in place, jumping jack, one-hand wave, two-hands wave, running, walking, and skipping on one leg. Sample images are shown in Figure 5.3. The extracted masks that come with this dataset make it easy to localize the human figures in each frame. For this dataset, most experiments have applied the leave-one-out cross-validation scheme for training and experiments. Our proposed method proved effective with a considerable rate of classification even with training done on one video and testing carried out across the remaining 9 videos.

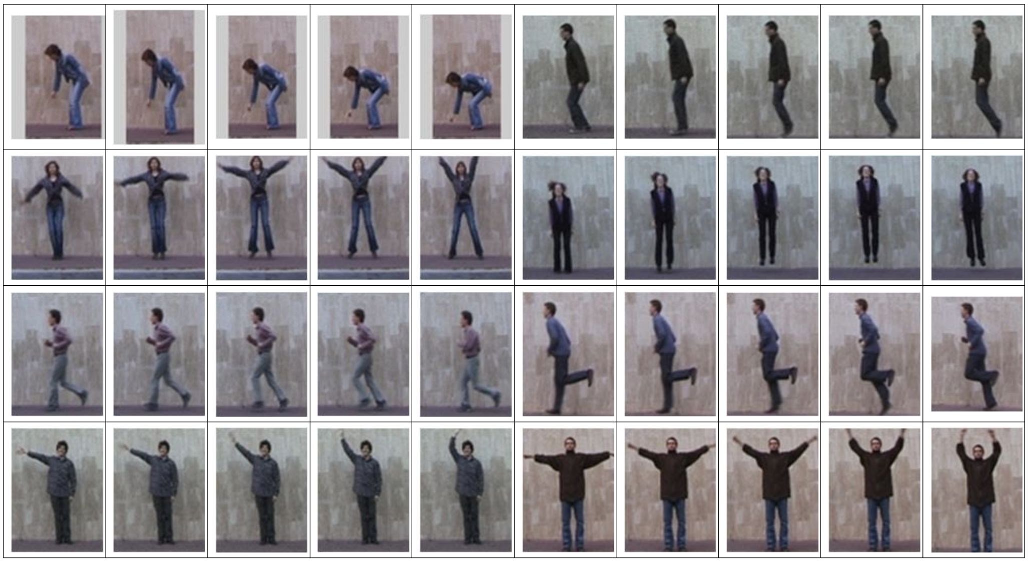


Figure 5.3 A snap shot of selected images from Weizmann dataset

5.1.3 Own-generated database (OGD)

This database contains our own generated video dataset with five subjects performing eight different actions of bending, walking, jump-jack, jump-place, one-hand wave, two-hands wave, jump-place wave, and running. The videos are taken at 15 frames per second. Each frame is of the size 320 x 240 with images of 24 bit RGB. The videos are also available in binary form as we have applied binarization for background removal. We show the RGB data images from OGD in Figure 5.4. The experimental results in [38] was achieved through OGD.

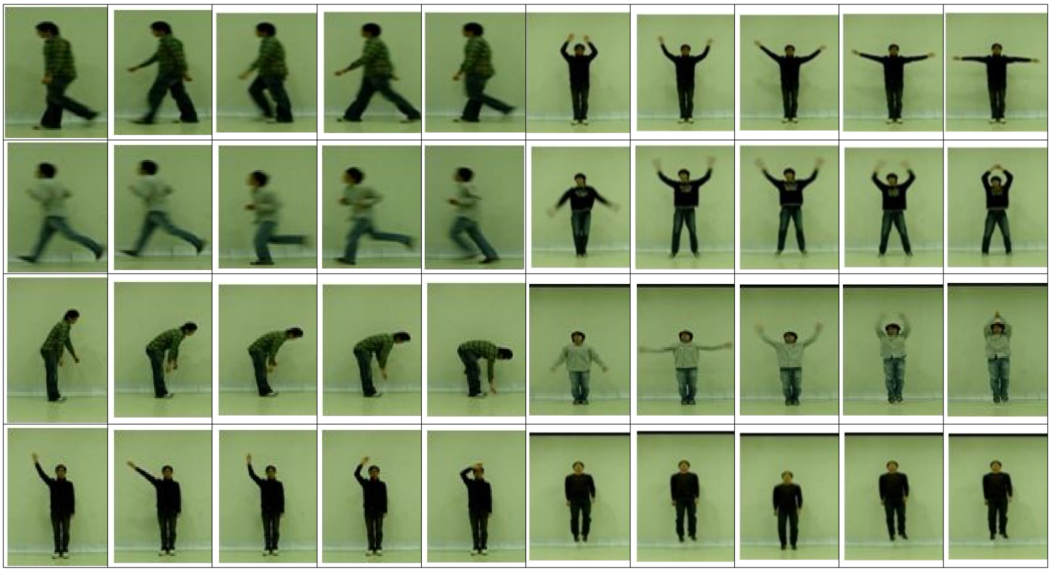


Figure 5.4 Raw data from our own-generated video database

5.2 Results and analysis

The descriptor reached has a log-polar of distance transformed images. Distance transformation of an image is defined as a map that assigns to each pixel the distance to the nearest black pixel. This forms a feature that can be used to distinguish between two or more images. Most applications use the distance transformation as a smoothening technique in the feature space. It is related to morphological dilation operation which we mentioned earlier in this work and is preferably used to explicitly searching for correspondences of features in images. The output is close to giving the edges of an image where the pixels are more pronounced and able to characterize image pattern. We illustrate the efficiency of distance transform on the summed up edge images and their distance transformed counterparts in Figure 5.5. Distance transform gives us a more simple but still powerful representation with enough information for discrimination purposes.

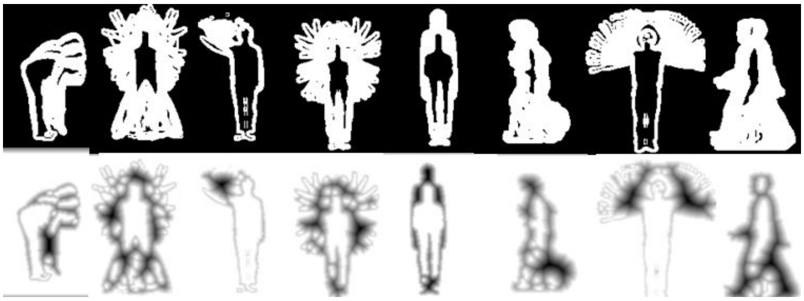


Figure 5.5 Silhouette edge models and their distance transform for 8 different actions from OGD.

The descriptor as we now know should bear salient features able to discriminate in a pool of images, particular traits. Log - polar transformations solves scale-invariance experienced between images. We have proved in our experiments that the log-polar images have the capacity to do this and wade off small perturbations and size disparity that may exist during training and testing of the model images. We show some of the log-polar counterparts of the models in Figure 5.6.

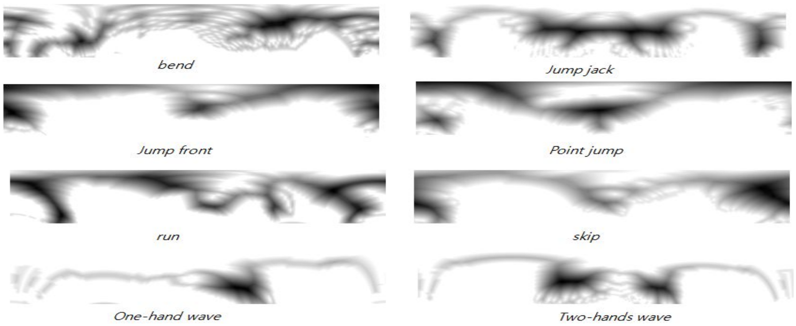


Figure 5.6 Log-polar transformations of the action models trained as labeled.

For matching between actions to happen, training for a database template image must go through the whole process as explained. The input image to be classified will also go through this process to model the action to conform to the specifics like size and form for comparison.

Table 5.1 shows some of the results of dissimilarity among various action model images by using Hausdorff distance measure. Hausdorff distance measure compared the action models by measuring how much difference exist within an action itself and even among other actions. Check for the explanation of Hausdorff distance as a dissimilarity measure. The experiment was done on Weizmann human action database.

Table 5.1 Maximum distance value of $h(A,B)$ and $h(B,A)$

	Bend	**Jack**	**Fjump**	**Pjump**	**Run**	**Skip**	**Wave1**	**Wave2**
Bend	0	107	107	135	107	95	132	95
Jack	107	0	55	132	67	55	132	73
Fjump	107	55	0	132	67	55	132	73
Pjump	135	132	132	0	135	132	132	126
Run	107	67	67	135	0	67	132	73
Skip	95	55	55	132	67	0	132	73
Wave1	132	132	132	132	132	132	0	126
Wave2	95	73	73	126	73	73	126	0

Figure 5.7 is an expression of Table 5.1 which shows sampled experiment on eight actions compared against the rest of the models. Where the bar is empty on the graph indicates that the action was compared against itself and express absolute similarity, so the value is, 0. Just like with all distance measurement techniques, the smaller the value in between actions, the closer the images in similarity.

Similar results can be observed in Figure 5.8 where the experiment was done on OGD with the model templates shown in Figure 5.6 used in comparison with the input query images. First, Table 5.2 indicates the Hausdorff distance value computed between each actions as indicated. From these values we have tabulated a general bar graph that capture each action compared to the rest of the others listed on the side list of the graph. A generalization of the performance of our descriptor is also given in Figure 5.9 which reveals the average differences among all actions tested. The line graph in Figure 5.9 (b) is given as a supplement to express clarity in the differences that exist by giving exact points.

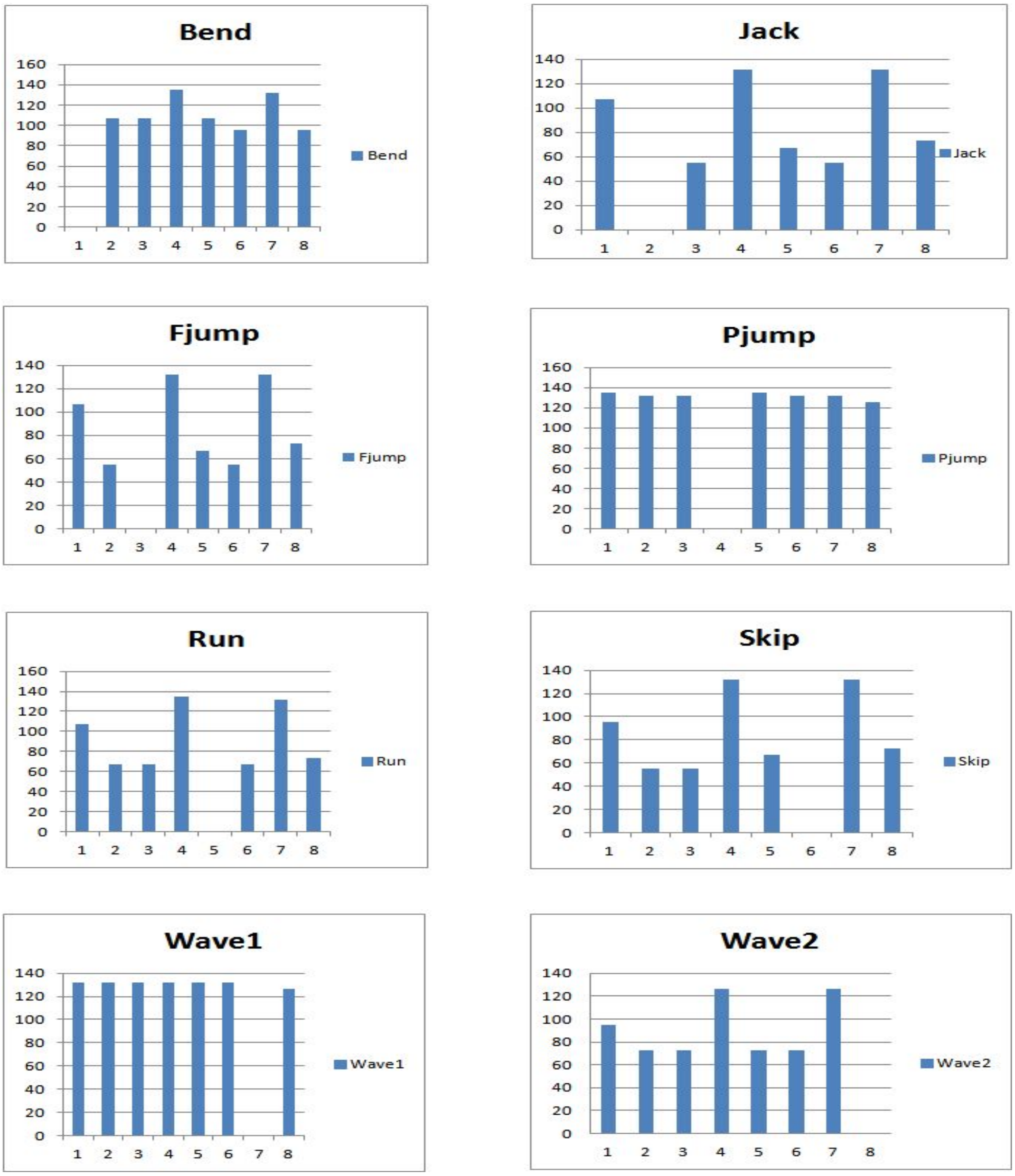


Figure 5.7 The sampled comparison graph data (Weizmann database). 1 to 8 represent ‘bend’, ‘jack’, ‘forward_jump(fjump)’, ‘point_jump(pjump)’, ‘run’, ‘skip’, ‘one-hand wave(wave1)’, ‘two-hands wave(wave2)’ actions, respectively.

Table 5.2 Minimax Hausdorff distance value on OGD

	Bend	JumpJ	WaveI	Pjump	PjumpJ	Run	WaveII	Walk
Bend	0	446	183	362	425	371	319	427
JumpJack	278	0	174	244	330	295	270	352
WaveI	497	506	0	475	508	497	322	508
PJump	346	462	180	0	433	360	295	446
PJumpJack	312	400	152	275	0	328	266	375
Run	341	446	173	329	413	0	299	412
WaveII	432	462	259	416	461	437	0	442
Walking	264	392	166	273	360	295	303	0

The graph shown in Figure 5.8 are expressions of the values in Table 5.2 comparing each action labeled with the rest of other actions. It can be further generalized to produce the following bar and line graphs in Figure 5.9 for clarity.

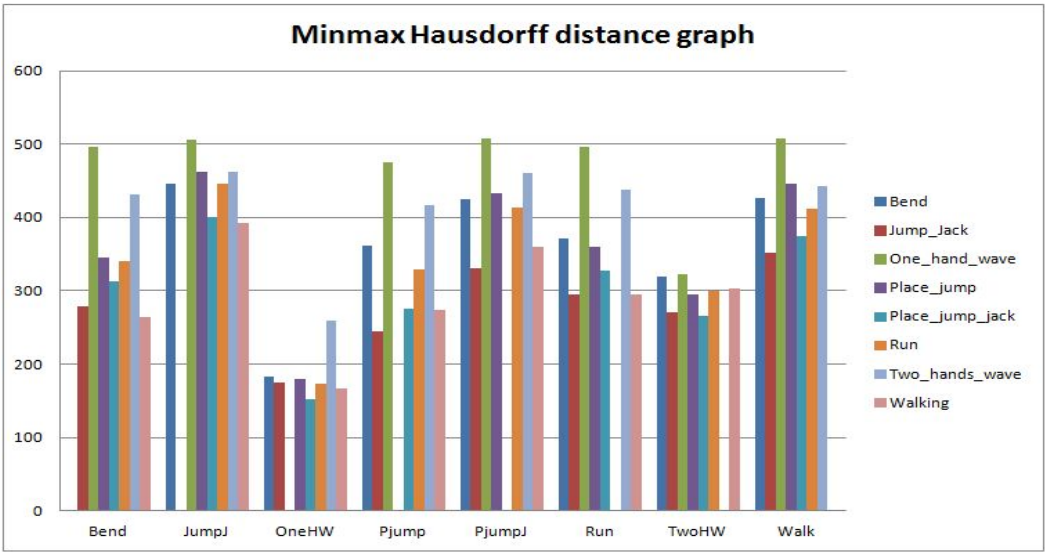
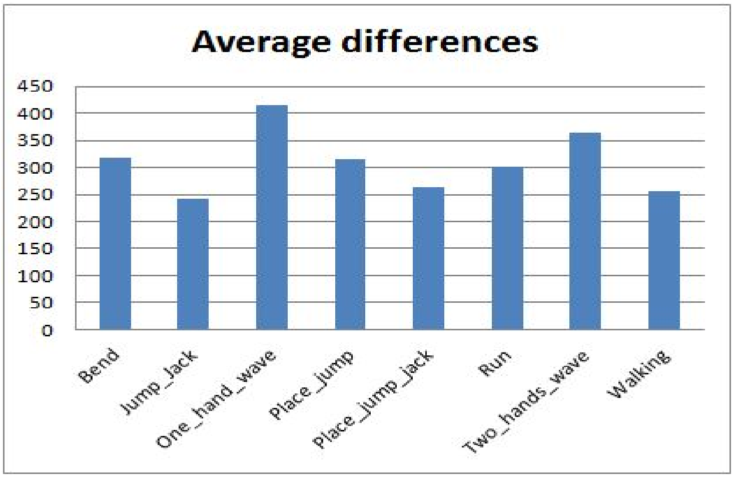


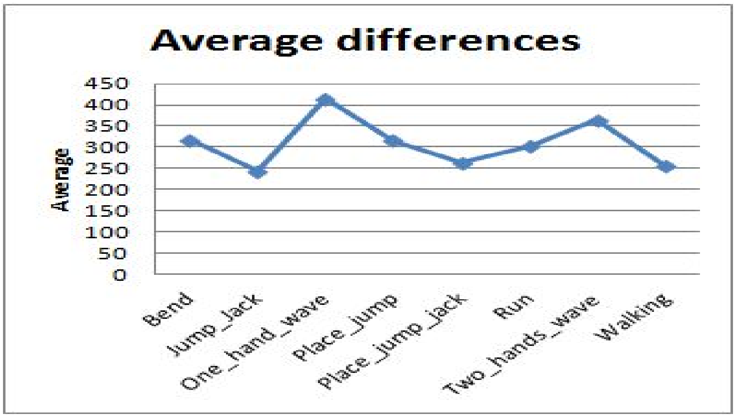
Figure 5.8 Minimax distance comparison among actions in OGD.

Figure 5.8 show a minimax distance graph among actions tested. We already stated how Hausdorff distance work in Chapter 3. There are two ways to use the computed distance values. Two images A and B are being compared both ways. A compared to B and vice versa. The distance in both cases are not the same in value but potray the same result if a set is used; that is we can decide to use the lowest distance computed or use the highest distance. In Figure 5.8, the average between the lower and the higher distance was used dubbing it the minimax distance graph. The results proved not so much difference as the trend read in both methods exhibited similarity between them.

Averages computed through the maximum and minimum values show a clear demarcation between different actions. Hausdorff distance value in some cases seem very close but visual representation of various descriptors developed show quite a margin that exist throughout the template models.



(a)



(b)

Figure 5.9 Average differences on OGD database by (a) bar graph (b) line graph

We can notice some very narrow gap between particular actions against each other. This disparity was realized in all the three databases used for testing the method. Each database had its own challenges that needed to be overcome. Actions like 'walk', 'jog', and skipping exhibited similarities between them across the databases.

Table 5.3 Sampled comparison of trained data between OGD and Weizmann databases using Hausdorff distance measure

Weizmann OGD	Bend	Jack	F_jump	P_jump	Run	Skip	Wave1	Wave2
Bend	3.5	100.8	150	117.2	83.9	51	84.5	183.5
J_jack	156.5	22	188.5	172	139	129.9	145.2	83.4
Wave1	90.6	16.2	154	139	70.4	92.9	4.2	28.6
P-Jump	48.8	21	66.9	0	12	2	28.7	14
Run	102.7	77	148	105.6	11	55.5	126.8	32.8
Wave2	151.2	111.4	201.4	197.8	152	139.5	166.2	70

Table 5.3 is a comparison of some of the actions we took from our own recorded video data (OGD) and one of the benchmark databases, Weizmann database. The recognition via Hausdorff distance performed fairly well with the distances as shown in the table. The lower the value, the closer the actions are to each other. The distance values between two inter-databases did not yield to zero frequently except for the 'point-jump' action. The recognition process followed the criteria explained in the proposed algorithm starting from object identification in videos to the template modeling. The results exhibited some closeness in particular actions but still the least-distance was clear. Some actions still presented similarity challenge, as can be witnessed in the experiment among 'point-jump', 'skip', and 'run' actions. The Hausdorff distance measure provided a value just enough to separate the actions and to let us claim the effectiveness of our proposed method. High classification performance is achieved, empirically.

A graph is shown in Figure 5.10 for clear difference that exist among several actions that we modeled and represented in log-polar form. In the graph, we can notice outright that the difference between 'two-hand wave' and 'jump-jack' actions compared to 'point-jump' action revealed biggest dissimilarity on average. From this particular experiment, 'run', and 'one-hand wave' actions can be seen quite close to each other.

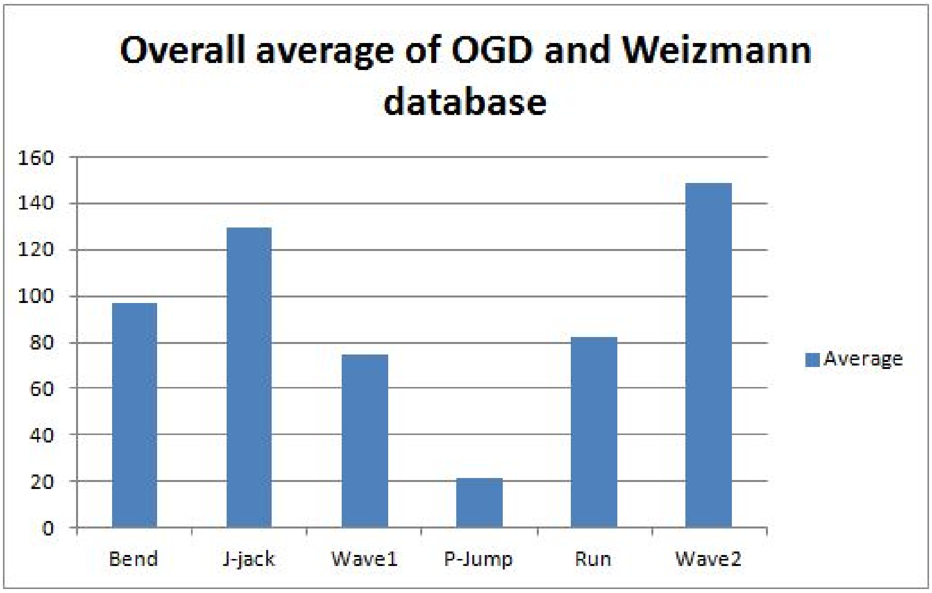


Figure 5.10 Overall average of six OGD-based dataset compared with 8-templates from Weizmann database

Continued analysis of our method was also carried out with Weizmann database where four individuals (Daria, Lena, Ido, and Shahar) [37] performing four different actions of 'bend', 'one-hand wave', 'two-hands wave', and 'jump-jack', was tested on. The four individuals are of different sizes and wore differently. The variations could be taken as a challenge to address invariant factors. The distance analysis on the final descriptors for all the actions trained matched well based on the minimum distance value computed through Hausdorff distance measure. The DT of the four mentioned individuals, the table showing the distance value analysis, and the graph originating from the table are shown in Figures 5.11 and 5.12 and

Table 5.4. An individual's Log-polar descriptor images are also shown. The visual dissimilarity of the patterns formed can be openly seen in the images through the human eyes.

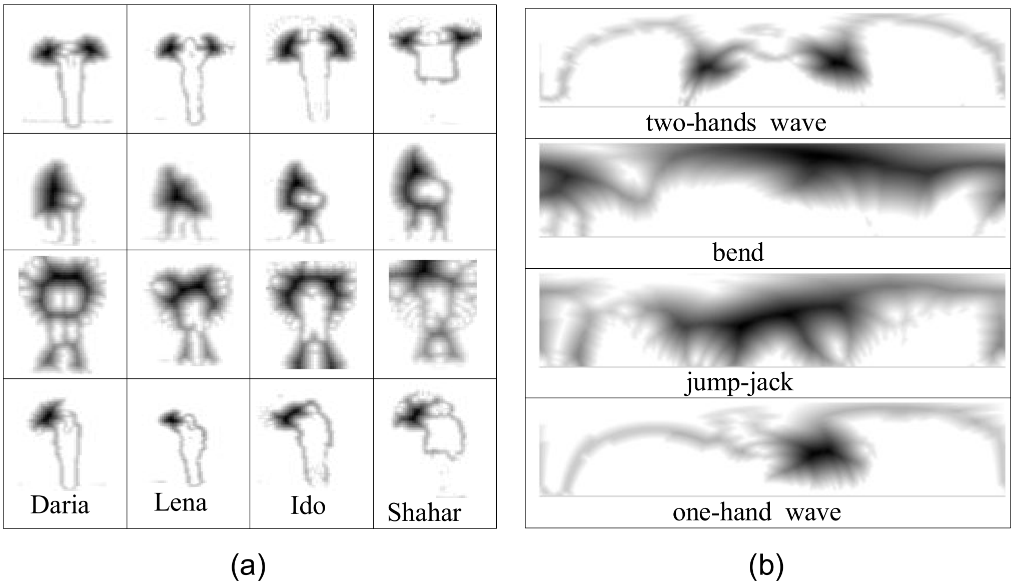


Figure 5.11 Weizmann database (a) The distance transform images of four individuals. (b) Log-polar transformations of the individual on first row (Daria)

Table 5.4 derive the distance values from Hausdorff computation of the images in Figure 5.11 (a). Next to it is the comparison graph expressing the separability between the performed activities.

Table 5.4 Distance measures from each other from Figure 5.11

Weizmann DB	**Bend**	**JJack**	**Wave1**	**Wave2**
Bend	311	352	140	150
JJack	455	270	160	205
Wave1	527	450	227	286
Wave2	520	448	315	217

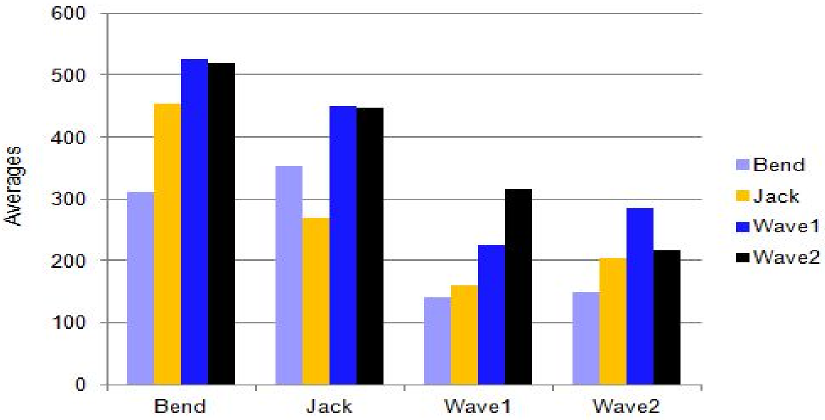


Figure 5.12 similarity between four actions performed by four different individuals

Log-polar imaging geometry is a biologically inspired approach to human vision where by it transforms the original two-dimensional captured image into a spatially variant (retina-like) representation [70].

From the graph in Figure 5.10, the use of the log-polar representation has empowered our method to perform even better. Inspired by a biological vision and data compression quality of the technique, log-polar images allow faster sampling rates on artificial vision systems and still maintains the size of the field of view without loss of important information. Log-polar images also increased the size range of objects that can be tracked using simple translation model. By applying log-polar transformation (LPT), data has been reduced through the sampling. As been stated, log-polar transform resembles the retina in structure. Image sampling reduces the resolution at the image periphery but leaves high resolution in the fovea (central) region to be dominant. Overall image resolution is lowered while a small region of interest is kept intact. This characteristic help improve the performance of feature tracking or identification, against a coarsely sampled background elements in the image periphery. Above all, scaling and rotations in original template image are turned to translations across the transform's axes. This facilitates matching tasks. And for this reason, we strongly believe that our method can perform robustly and better within the areas suggested.

The validation technique that we chose for the performance of our descriptor and classification exhibited minimal error rate. Cross-validation methods are a whole class of model evaluation that require training of part of the dataset and not the entire dataset for the learner. Holdout method of cross-validation splits the dataset into two groups; training set, used to train the classifier and; test set, used to estimate the error rate of the trained classifier.

This work has used K-fold cross-validation, an improvement of holdout method. The databases we have used have limited number of classes of actions, which let us divide the dataset in 8 subsets. The holdout method was repeated eight times according to the number of subsets. The eight subsets are used alternately as a test set and the other forms as a training set. The method is appropriate because each data in the data set gets to be in the training dataset and also the testing dataset. The procedure is enhanced by our classification method which uses Hausdorff distance measure. Recall that Hausdorff distance measure is executed to record both the minimum and maximum distances between images *A* and *B* and vice versa, as in Equations (4.1) and (4.2). For each experiment in K-fold cross-validation, we used K-folds for training and remaining ones for testing. Testing was also done between databases containing different actors with the same actions. The error is estimated as the average error rate as in Equation 5.1.

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i \qquad (5.1)$$

where E is the true error, K - the total number of experiments carried out, and i is an index that stands for whichever data it is associated with.

Another widely used evaluation method is a Leave-one-out cross-validation (LOO). It is considered an extreme form of K-fold cross-validation in which the whole data in a dataset except one form the training set. For a dataset with N examples, the requirement is that N number of experiments has to be performed. For each of the experiments performed, use N-1 examples for training and the remaining single data is used as a test example. The estimated error is computed in the same way as in Equation 5.1.

Re-substitution is another method which uses all the available data in a dataset for both training and testing for validation. The method executes perfectly with the trained data but has over-fitting and could perform very poorly with a testing carried out on unseen data.

In search for a better and simple similarity measure, we compared the performance of Euclidean distance [73] on the same dataset. This is one of the most commonly used image metrics. Basically, given two images to be compared, M and N, $x = (x_1, x_2, ..., x_{MN}), y = (y_1, y_2, ..., y_{MN})$, where x_i, y_i represents gray levels at locations within the given images. The Euclidean distance $d_E(M, N)$ is given by Equation (5.2).

$$d_E^2(M, N) = \sum_{i=1}^{MN} (x_i - y_i)^2 \qquad (5.2)$$

The experiments in Figure 5.11 compared two databases of Weizmann and OGD with the more challenging KTH database. To state it again, KTH presents homogeneous background, scale variation, different clothing and illumination variation, providing a good platform to test our descriptor under constraint environment. The descriptor produced good matching results with both Hausdorff and Euclidean distance measures. KTH contain six different actions of which we used three common ones found in both OGD and Weizmann databases as in Table 5.5. These are walking, running, and two-hands waving actions.

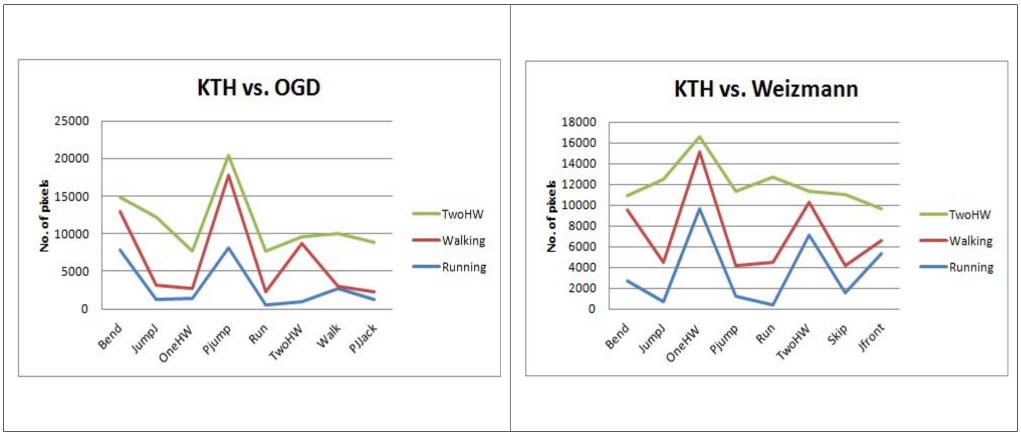
Table 5.5. Euclidean distance between KTH against (a) OGD (b) Weizmann databases

ODG KTH	Bend	JumpJ	OneHW	PJump	Run	TwoHW	Walk	PJJack
Running	7788	1183	1455	8080	516	1022	2690	1297
Walking	5219	2033	1199	9682	1779	7675	319	950
TwoHW	1910	8956	4967	2663	5371	894	6986	6623

(a)

Weizmann KTH	Bend	JumpJ	OneHW	PJump	Run	TwoHW	Skip	JFront
Running	2675	727	9601	1200	387	7126	1571	5324
Walking	6888	3791	5554	2965	4044	3107	2637	1289
TwoHW	1356	7969	1426	7180	8273	1089	6843	3057

(b)



(a) (b)

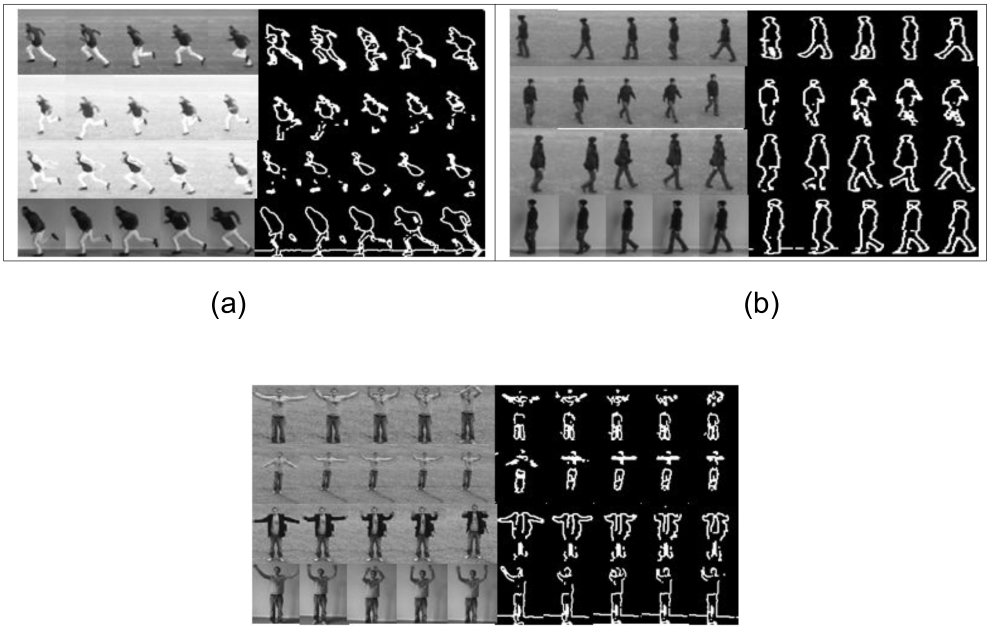
Figure 5.13 Euclidean distance graph (a) KTH against OGD, (b) KTH against Weizmann database

The higher the Euclidean distance value between the images being compared the wider the dissimilarity. Euclidean distance uses the pixel counts which is very important for images which harbour spatial relationships between pixels. The descriptor used captures actions with high intensity exhibited around the region of interest; that is, where much movements have been realized. This makes mapping of images onto each other easier during recognition directly and pixel-wise comparison viable.

5.3 Challenges

KTH human action database which presented most challenges contains sequences of six classes of actions performed by 25 subjects in four different conditions, S1 to S4. According to [39], the symbols S1 to S4 represents actions on homogeneous backgrounds, plus scale variations, different clothing, and lighting variations, respectively. The benchmark databases used do not carter for occluded images making our method to be vulnerable to occlusion. With the variations in scale , clothes, and lighting, KTH database presents different environmental conditions under which we can test our method. For localization of the actor (human figure), a bounding box is used. External interference have introduced noise in the images which should be removed before edges are detected. This is achieved through median filtering. Sobel edge detection is used to detect the mage edges and then convolution is done by convolution between the image and a formed mask. For visual clarity, morphological operations are used to improve edge quality. The experiments on this database was more taxing, especially cross-validation of actions from different databases.

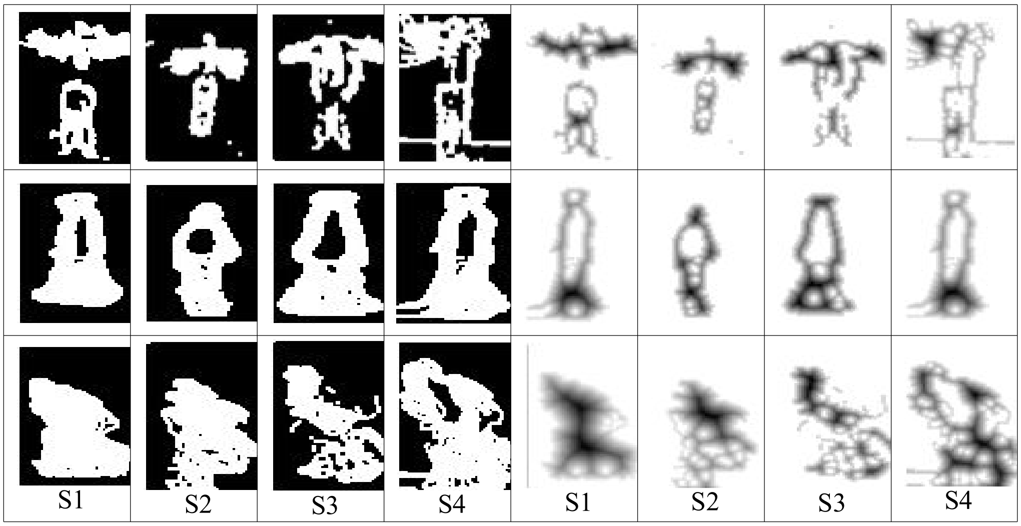
Figure 5.12 shows some results of image edges affected by the mentioned factors like illumination and clothing among others. These factors reduced performance in representing the actions thereby affecting classification.
In Figure 5.14 (a), the two conditions of scale and change of clothing with a bright lighting caused our model to be deformed in a way that could not expressly define running action in one instant. The same was experienced in (c) where the clothing and the background looked almost similar. This presented difficulty in edge extraction from the targeted foreground objects.



(a)

(b)

(c)



(d)

Figure 5.14 KTH database showing (a) running, (b) walking, and (c) two-hands wave under four different environments. (d) models and their distance transforms. In (a)-(c), first row down to the fourth row represent S1 to S4, respectively.

In overall, the proposed method for the descriptor was able to represent the actions with the limited edges closer to the original representations carried out in an ideal environment. These representations are shown in Figure 5.14 (d) above.

There are many other descriptors that can be used depending on the features to be extracted. Some examples of descriptors include scale-invariant feature transform (SIFT) [74], gradient location and orientation histogram (GLOH) [75], principle component analysis – SIFT (PCA-SIFT) [76], differential invariants [77], shape context [42], spin images [78], and moment invariants [79] to mention but a few away from the log-polar descriptor which we have used.

The problem we are trying to solve here is the similarity between image descriptors and the extracted video features, to which descriptors provide solutions. Two or more images to be compared could be misaligned due to rotation, scale, shear and translation. We have used log-polar transformation to accommodate arbitrary rotation angles and a wide range of scale changes expected to occur in probe and gallery images. Direct comparison needs some kind of normalization for its correctness to be ascertained. Descriptors achieve this kind of task. We have not found work that is same with our work in human action descriptor recognition. However, log-polar transform is applauded in various researches for excellent alignment of perturbed images and is used as a descriptor in [80 81 82]. The many successful applications and the good review the method has received made it find its way in our work. Through experiments, we have proved the same and shown that the representation through log-polar transformed images is compact. The evaluation criterion of the recognition process was based on the recall-precision method whereby the number of correct and false matches between two different action images was computed for the classification rate.

Chapter 6

Conclusions

We have proposed a powerful image descriptor for representing and recognizing human actions in videos. The method based on silhouette boundary edges , with a compact human action descriptor through log-polar representation proved sufficient and efficient enough during recognition. For salient feature extraction from video data with limited redundancy of the human pose, information entropy was applied for key frame selection process. Frame-blocking is used to calculate the spread of pixels within the bounding box. A unique representation of various actions was formed through silhouette boundary edges stacking. The resultant pattern is enhanced through distance transform of the summed edges then transforming them into log-polar templates. A well known and researched Hausdorff distance is used as a similarity measurement to classify the actions into appropriate groups. Same experiments carried out with the traditional Euclidean distance was not far from the results produced by Hausdorff distance measure. Validation of our process has been done on three databases; KTH database, Weizmann database, and our own-generated database (OGD); the first two being benchmark databases for most researchers and their work. Our experimental analysis prove that the templates performed exceptionally well for the purpose they were intended to execute. The application of our proposed method is very action-specific with the constrained environment as per the databases conditions. We dealt with particular actions in very particular environments. Even though a database like KTH presented challenges like distance, illumination, and scale variance, which made our descriptors look a little skewed and sketchy in a way, the summation of extracted boundary edges formed a visible model that could be processed give reasonable descriptors. The experimental results with this database presented great challenge. In general, according to the data given for training and the actions performed, our proposed method is considered a success. The performance of the proposed method is very high on the benchmark dataset, Weizmann. However, more research need to be done for a robust descriptor of various activities under varying environments to be properly represented. Our method cannot be used where there are multiple subjects carrying out different activities concurrently. Occlusion and individual object segmentation could be an uphill task in such situations. A dynamic 3-D feature extraction from complex scenes seem viable for this kind of work and remain to be our future research area. Furthermore, a different method of classification will have to be explored for the independent activities residing within the same video frame to be identified and separated from the rest of the frame.

References

[1] Agarwal. A. and Triggs. B, "3D human pose from silhouettes by relevance vector regression", Computer Vision and Pattern Recognition, 2004. CVPR 2004.

[2] Fenjun. L.V and Nevatia. R, "Single View Human Action Recognition using Key Pose Matching and Viteri Path Searching", Computer Vision and Pattern Recognition, 2007. CVPR June 2007. pp 1 - 8.

[3] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, “Behavior recognition via sparse spatio-temporal features”, International Workshop on Performance Evaluation of Tracking and Surveillance, 2005, pp. 65 – 72.

[4] Aaron F. Bobick and James W. Davis, "The recognition of Human Movement Using Temporal Templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001.

[5] Ali. A, Aggarwal. J.K, "Segmentation and recognition of continuous human activity", Detection and Recognition of Events in Video, 2001. IEEE Proceedings.

[6] S. Eickeler, A. Kosmala, G. Rigoll, "Hidden Moarkov model based continuous online gesture recognition", Proceedings Fourteenth International Conference on Pattern Recognition, Volume. 2. IEEE Computer Society, pp. 1206 - 1208.

[7] Elgammal. A, Shet. V, Yacoob. Y, Davis. L. S, "Learning dynamics for

exemplar-based gesture recognition", Computer Vision and Pattern Recognition, 2003, Proceedings. 2003.

[8] Daniel Weiland, Remi Ronfard, Edmond Boyer, "Free viewpoint action recognition using motion history volumes", Computer Vision and Image Understanding, Volume. 104, 2006. pp. 249 -257.

[9] Lena Gorelic, Moshe Blank, Eli Shechtman, Michal Irani, Ronen Basri, "Actions as space-time shapes", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (12) (2007) 2247 - 2253.

[10] Daniel Weiland, Remi Ronfard, Edmond Boyer "A survey of vision-based methods for action representation, segmentation and recognition" Computer Vision and Image Understanding 115, 2(2011) 224 – 241.

[11] G. Johansson, "Visual Perception of Biological Motion and Model for its Analysis", Perception and Psychophysics, 14(2): 210-211, 1973.

[12] Cedras. C, Shah. M, "A survey of motion analysis from moving light displays", Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference. pp. 214 – 221.

[13] Daniel Weiland, Remi Ronfard, Edmond Boyer,“A survey of vision-based methods for action representation, segmentation and recognition”, Computer Vision and Image Undertsanding (2010). Volume.115, Issue:2. pp 224 – 241.

[14] W. Forstner, E. Gulch, “A fast operator for detection and precise location of distinct points, corners and centres of circular features”,

Intercommission Conference on Fast Processing of Photogrammetric Data, 1987, pp. 281 – 305.

[15] C. Harris, M. Stephens, “A combined corner and edge detector”, Alvey Conference, 1988, pp. 147 – 152.

[16] D. G. Lowe,“Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision 60 (2) (2004) 91 – 110.

[17] I. Laptev, T. Lindeberg,“Space-time interest points”, IEEE International Conference on Computer Vision, 2003, pp. 432 – 439 vol.1.

[18] Davis J. W, "Apperance-based motion recognition of human actions", M.I.T. Media lab Perceptual Computing Group Tech. Report Number 387. 1996. pp. 51.

[19] Bobick, A., Davis, J. "An appearance-based representation of action", International Conference on Pattern Recognition, 1996. pp. 307 - 312.

[20] Wang L.,Hu W., Tan T.,"Recent development in human motion analysis", Pattern Recognition Volume 36, 2003. pp. 585 - 601.

[21] Bruce D. Lucas, Takeo Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", International Joint Conference on Artificial Intelligence pages 674 - 679, 1981.

[22] Yin, Z., Collins, R., "Moving object localization in thermal imagery by forward-backward MHI", Proceedings IEEE Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum. NY, pp. 133 -

140, June 2006.

[23] Yau, W., Kumar, D., Arjunan, S.,"Visual speech recognition using image moments and multiresolution wavelet", Proceedings Conference on Computer Graphics, Imaging and Visualization, pp. 194 - 199, 2006.

[24] Bradski, G., Davis, J.,"Motion segmentation and pose recognition with motion history gradients", Proceedings IEEE Workshop on Applications of Computer Vision. pp. 174 - 184, December 2000.

[25] Haritaoglu, I., Harwood, D., Davis, L. S., "W4: real-time surveillance of people and their activities", IEEE Transcations PAMI 22(8), pp. 809 - 830, 2000.

[26] Sun, H.Z., Feng, T., Tan, T.N.,"Robust extraction of moving objects from image sequences", Proceedings Asian Conference on Computer Vision, pp. 961 – 964, 2000.

[27] Collins, R.T., Lipton, A., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L. "A system for video surveillance and monitoring VSAM final report", CMU-RI-TR-00-12, Technical Report, Carnegie Mellon University, pp. 69, 2000.

[28] Kameda, Y., Minoh, M."A human motion estimation method using 3-successive video frames" Proceeedings International Conference on Virtual Systems and Multimedia, pp. 6, 1996.

[29] Rafael C. Gonzalez, Richard E. Woods,"Digital Image processing,

Second Edition", Prentice Hall, 2001. pp. 220 - 243.

[30] Kalman, R. E. "A new approach to linea filtering and prediction problems" Transaction of the ASME - Journal of basic Engineering, 82(series D). pages. 35 - 45, 1960.

[31] Ivan L., Barbara C., Christian S., Tony L. "Local velocity-adapted motion events for spatio-temporal recognition" Computer Vision and Image Understanding, (CVIU) 108(3), 2007. pp. 207 - 229.

[32] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, Benjamin Rozenfeld, "Learning realistic human actions from movies", Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008. pp. 1 - 8.

[33] Marcin Marszalek, Ivan Laptev, Cordelia Schmid,"Actions in Context", Proceedings on Conferences on Computer Vision and Pattern Recognition (CVPR'09) Miami, FL, June 2009, pp. 1 - 8.

[34] Mikel D. Rodriguez, Javed Ahmed, Mubarak Shah, “Action MACH: a spatio-temporal maximum average correlation height filter for action recognition", Proceedings on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1 - 8.

[35] Daniel Weiland, Remi Ronfard, Edmond Boyer, "Free viewpoint action recognition using motion history volumes", Computer Vision and Image Understanding (CVIU) 104(2-3) (2006) 249 - 257.

[36] Christian Schuldt, Ivan Laptev, Barbara Caputo, "Recognizing human

actions: a local SVM approach", Proceedings of the International Conference on Pattern Recognition (ICPR'04), 2004. vol.3, Cambridge, United kingdom, 2004. pp. 32 - 36.

[37] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, Ronen Basri, "Actions as space-time shapes", Proceedings of International Conference on Computer Vision (ICCV'05), vol.2, Beijing, China, October 2005, pp. 1395 – 1402.

[38] Geum-Boon Lee, Wilfred O. Odoyo, Jeong-Nam Yeom, Beom-Joon Cho, "Extraction of key postures using shape contexts", Proceedings of the 11th International Conference on Advanced Communication Technology (ICACT'09), February 2009, pp. 1311 - 1314.

[39] Keren, D., Osadchy, M., Gotsman, C., "Antifaces: Anovel, fast method for image detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(7). pp. 747 - 761.

[40] C. E. Shannon,"A mathematical theory of communication", Bell System Technical Journal, vol 27, pp. 379 - 423 and 623 - 656, July and October, 1948.

[41] Richard O. Duda, Peter E. Hart, David G. Stork, " Pattern Classification, Second Edition, 2001, Appendix A. pp. 630 - 633.

[42] S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape context", IEEE Transactions on Pattern Analysis and machine Intelligence, vol.24. No. 24, pp. 509 - 522, 2002.

[43] Yan Chen, Qaing Wu, Xiangjian He, Chnhua Du, Jie Yang, "Extracting key postures in a human action video sequence", Multimedia Signal Processing, 2008 IEEE 10th Workshop, October 2008, pp. 569 - 573.

[44] Alasdair Mc'Andrew, " Introduction to Digital Image Processing with MATLAB" 2004 Course Technology, Chapter 10, pp. 261 - 296.

[45] B. Vanajakshi, B. Sujatha, K. Srirama, "A study on implementation of advanced morphological operations", IJCSNS International Journal of Computer Science and Network Security, vol. 10 No. 3, March 2010, pp. 6 - 9.

[46] M. G. Ullah, B. S. Chowdhary, Shiraz Latif, A. Qadeer Rajput, Javed Ahmed, "Pattern matching algorithm using polar spectrum in retina recognition for human identification system", Australian Journal of Basic and Applied Sciences, 5(10): 1385 - 1392, 2011.

[47] V. Javier Traver, Filiberto Pla, " Log-polar mapping template design: Fromtask-level requirements to geometry parameters", Image and Vision Computing 26 (2008) 58 - 74.

[48] Siavash Zokai and George Wolberg, "Image registration using Log-polar mappings for recovery of large-scale similarity and projective transformations", IEEE Transactions on Image Processing, vol. 14, No. 10, October 2005, pp. 1422 - 1434.

[49] Rittavee Matungka, Yuan F. Zheng, Robert L. Ewing, " Image registration using adaptive polar transform", IEEE Transactions on

Image Processing, vol. 18. No.10, October 2009, pp. 2340 – 2354.

[50] Meng H., Pears N., Bailey C., "Recognizing human actions based on motion information and SVM", Proceedings of IEEE International Conference for Intelligent Environments, 2006. pp. 239 - 245.

[51] Schuldt C., Laptev I., Caputo B., "Recognizing human actions: a local SVM approach", Pattern Recognition, 2004, ICPR 2004, Proceedings of the 17th International Conference, August 2004, vol. 3. pp. 32 - 36.

[52] Sarvaiya J. N., Patnaik S. Bombaywala S., "Image registration by template matching using normalized cross-correlation", Advances in Computing, Control and Telecommunication Technologies, 2009. ACT'09 December. pp. 819 - 822.

[53] D. Huttenlocher, G. Klanderman, W. Rucklidge,"Comparing images using the Hausdorff Distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15. No. 9, pp. 850 - 863, September 1993.

[54] M. P. Dubuisson and A. K. Jain,"A modified Hausdorff distance for object matching," Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, October 1994, pp. 566 - 568.

[55] Dong-Gyu Sim, Oh-Kyu Kwon, Rae-Hong Park,"Object matching algorithms using robust Hausdorff distance measures," IEEE Transactions on Image Processing, vol.8, No. 3, March 1999.

[56] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques

for image matching,"Proceedings of 5th International Joint Conferences on Artificail Intelligence, pages 659 - 663, 1977.

[57] G. Borgefors, "Hierarchical chamfer matching: A parametric edge matching algorithm," IEEE Transactions on Pattern Analysis and Machine Intelligence., 10(6):849 - 865, November 1988.

[58] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing (Second Edition). Upper Saddle River, New Jersey. Prentice-Hall, Inc. 2002. pp. 134 – 141.

[59] Alasdair McAndrew, "Introduction to Digital Image Processing with MATLAB," Thomson Learning, Inc. Course Technology, 2004. pp. 229 – 236.

[60] Nazli Ikzler, Pinar Duygulu, “Histogram of oriented rectangles: A new pose descriptor for human action recognition,” Image and Vision Computing 27 (2009)1515 – 1526.

[61] Ahmad, M., Lee, S.-W.,“ Recognizing human actions based on silhouette energy image and global motion description,” Proceedings of the IEEE Automatic Face and Gesture Recognition, pp. 523-588 (2008).

[62] Chen, D., Yang, L., “Exploiting high dimensional video features using layered Gaussian mixture models,” Proceedings of IEEE ICPR, pp.4 (2006).

[63] Meng, H., Pears, N., Bailey, C., “ Human action classification using SVM_2K classifier on motion features, LNCS Multi-media Content

Representation Classification and Security, vol. 4105/2006. pp. 458-465 (2006).

[64] Ahmad, M., Hossain, M.Z., “ SEI and SHI representations for human movemen recognition,” Proceedings of the International Conference on Computer and Information Technology (ICCIT), pp. 521-526 (2008).

[65] Bradski, G., Davis, J., “Motion segmentation and pose recognition with motion history gradients.” Machine and Vision Application vol. 13(3), pp. 174-184 (2002).

[66] Davis, J. W.,“Appearance-based motion recognition of human actions,” M.I.T Media Lab Perceptual Computing Group Technology. Report No. 387, pp. 51(1996)

[67] Ahad, Md. A. R., Tan, J. K., Kim, H., Ishikawa, S.,“Temporal motion recognition and segmentation approach,” International journal of Imaging Systems and Technology. vol. 9 pp. 91-99 (2009).

[68] Ahad, Md. A. R., Tan, J. K., Kim, H., Ishikawa, S.,“Analysis of motion self-occlusion problem due to motion overwriting for human activity recognition,” Journal of Multimedia. vol. 5(1), pp. 36-46 (2009).

[69] Babu, R., Ramakrishnan, K.,“Recognition of human actions using motion history information extracted from the compressed video,” Image Vision and Computing. vol. 22, pp. 597-607 (2004).

[70] Orrite, C., Martinez, F., Herrero, E., Ragheb, H., Velastin, S., “Independent viewpoint silhouette-based human action modeling and

recognition," Proceedings ofInternational Workshop onMachine Learning for Vision-based Motion Analysis (MLVMA'08) with ECCV, pp. 1-12 (2008).

[71] Wai Kit Wong, Chee Wee Choo, Chu Kiong Loo, and Joo Peng, "The FPGA Implementation of Log-polar Mapping," 15th International Conference on Mechatronics and Machine Vision in Practice (M2VIP08), 2-4 December 2008, Auckland, New-Zealand.

[72] Richard Alan Peters II, "On the computation of the discrete log-polar transforms,"unpublished, 22 April 2007.

[73] Liwei Wang, Yan Zhang, Jufu Feng, " On the Euclidean Distance of Images," IEEE Transactions on Pattern Analysis and Machine Intelligence. August 2005. vol 27 no. 8. pp 1334 – 1339.

[74] D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 2(60):91 – 110, 2004.

[75] Krystian Mikolajck and Cordelia Schmid, "A performance evaluation of local descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, 10, 27, pages 1615 – 1630, 2005.

[76] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," In Proceedings of the Conference on Computer Vision and Pattern Recognition, Washington, USA, pages 511 – 517, 2004.

[77] J. Koenderink and A. Van Doom, "Representation of local geometry in

the visual system," Biological Cybernetics, 55: 367 – 375, 1987.

[78] S. Lazebnik, C. Schmid, and J. Ponce, "Sparse texture representation using affine-invariant neighborhoods," In Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Winconsin, USA, pages 319 – 324, 2003.

[79] L. Van Gool, T. Moons, and D. Ungureanu, "Affine / photometric invariants for planar intensity patterns," In Proceedings of the 4th European Conference on Computer Vision, Cambridge, UK, pages 642 – 651, 1996.

[80] Wolberg, G. "Robust image registration using log-polar transformation," International Conference on Image Processing, vol 1, pages 493 – 496, 2000.

[81] Qi Wang, Ke Zhang, Youyi Jiang, Xianze Xiong, "The discrete algorithm of log-polar transformation," 1st International Symposium on Systems and Control in Aerospace and Astronautics, 2006, ISSCAA, 4 page 692.

[82] Yue Bao, Bin Qi, Fei Gu, " Facial recognition using partial log-polar transformation," IEEE / SICE International Symposium on System Integration (SII), 2011, pages 74 –77.

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Professor Beom-Joon Cho for his supervision and guidance throughout my course. I greatly express gratitude to my Committee Chair, Professor Yun-Bae Lee, and to the other committee members, Professors Yong_Geun Bae, Sang-Woong Lee, and In-Kyu Moon. Your input and advice during thesis presentations and discussions were of great help. Thanks for your contributions to this work.

Thanks to all the AI&PR Lab members for their daily support, especially from my senior, Dr. Geum-Boon Lee.

Finally, big thanks to my family members, my mom, and everyone else for standing by me through the many years spent away from them. It has been a path to nobility and I have appreciated. Thank you for the education and the support.

I dedicate this to you all.