CHAPTER 1

Introduction

In recent years, human action recognition due to its potential use in active computing in various research areas has aroused considerable interest. Such automated surveillance systems, human-computer interaction, smart home health care system and the regulation of human action recognition from video sequences such as problem-free gaming systems able to identify many important applications in different human tasks in the project addressed a reliable system, . The purpose of the input video sequences, kicking, etc., such as bending, bowling, boxing, jogging, jumping as low-level tasks are to develop an algorithm that can recognize. In depth view of both RGB video sequences as well as the human action recognition (HAR) was analyzed.

Interest in the subject many applications, both offline and online, is inspired by the promise. Automatic annotation of video news footage or music video example football matches in specific dance moves, handshakes find the catch, and enables more efficient search. Online processing, for example in traffic locations, but also to support aging in place in homes and houses for the elderly and children, automatic monitoring allows. Interactive applications, human-computer interaction or game, for example, automatically benefit from progress in the field of human action recognition.

Often, immediate action in time is assumed to be fragmented, but it might not always be practical. Recent work on the recognition of action is to resolve the issue. In addition, the performance of an action may be considerable variation in the rate. The action is recorded, the rate at which the motion features are used, especially when an action has a significant impact on the temporary limit. A robust human action recognition algorithm is invariant to view angle, light condition, different rates of execution.

Model-free approach, the features extracted from image sequences is generally employed as a descriptor of the action. Commonly used descriptors are silhouette, contour, optical flow, trajectory, or the silhouette and contour points of interest mostly is adopted, which is derived from, among other features. Create a series of 2D

contours of the human object spatiotemporal volume (STV) Yilmaz et al. [1] is presented. [2]. STV calculated by analyzing the geometric doing geometrical and structural analysis recognition.

To ensure human activity variations approach, Ben-Arie et al. [3] representation of the human body motion vectors that pose temporal sequences as described in these operations is irreversible. They are building a database of major body parts; Currency in which the activity templates were stored as entries in multidimensional hash tables. Voting multidimensional indexing and matching approach to improve the efficiency and stability were used in the validation phase. Recently, Lu in a video sequence automatically track multiple hockey players and simultaneously recognize their actions have developed a system. Hue, Saturation Values, and HOG descriptor to represent the color histogram information of image size, respectively, for baseball players were used. The HOG descriptor for the X and Y direction and magnitude of gradients based on both their learning used a 3-D histogram. Thus, their method is invariant to variations approach. The battle thus leant templates and can be updated from the training dataset. For action, Mltinomiyl a sparse logistic regression (SMLR) classifier action categories to classify your hog descriptors can be used.

However, widespread deployment of CCD cameras is more expensive and ineffective in human supervision. To automatically detect the shocking events rather than passively to record continuously, second generation surveillance systems were developed. Developed by the MIT Media Laboratory and the University of Maryland in the early years Pfinder [4] that this kind of system. The important feature of these systems robust detection, tracking, and classification algorithms lies in the ability to provide. Pfinder and W4 addition, much second-generation surveillance system recently emerged into existence. For example, the Defense Advanced Research Projection Agency (DARPA) project sponsored by HID, detection and identification of significant standoff distance [5] to identify the man in a human identification system linked to biometric technologies. Biometric technology, including terrorists, criminals, and other human-based threats to the Force Protection and Homeland Defense can provide useful early warning support, thus enabling humans to help faster and more accurate identification, and can. Differently, Information Society

Technologies (IST) project funded by the caviar, visual, used various information, including function, and hierarchal visual processes [6] through local images relevant knowledge to provide rich details objection. Information night crime detection and classification of customers in commercial behavior to their task can enable caviar. The video stream is dry to normal activities and activities in order to filter the UK's Engineering and Physical Sciences Research Council, the University of Edinburgh [7] project funded by the act. Detect and discriminate between similar interactions can track individuals using dynamic hidden Markov models. In addition, global probabilistic model images have been obtained in a short time, in view of the crowd during the tracking of persons was adopted to resolve the disparity.

Database Used and its Description: KTH database is used for training, classification, and detection. KTH has 6 activities and 4 scenarios and 25 actors. This all gives us a good set of action videos and this number turn out to be 100.

Challenges: To detect simple action of KTH dataset as Running, Hand-waving, Handclapping, Jogging, Walking and Boxing. Since the method is universal and works well on all the other dataset which represent simple activities.

Keywords: Grid, Cells, Spatio-Temporal, LDA, PCA, Nearest Neighbor algorithm, Support Vector Machine etc.

1.1 Motivation

Interest in the subject many applications, both offline and online, is inspired by the promise. Automatic annotation of video news footage or music video example football matches in specific dance moves, handshakes find the catch, and enables more efficient search. Online processing, for example in smart homes, but also to support aging in place in Homes, automatic monitoring allows. Interactive applications, human-computer interaction or game, for example, automatically benefit from progress in the field of human action recognition.

In addition, anthropometric differences between individuals, similar observations, especially non-cyclical action or actions that are adapted to the environment (walking, or by pointing to a certain place to avoid such obstacles), can be used for other tasks.

Overlap between the classes will be more action as the increasing number of classes, it would be more challenging. In some domains, the class label distribution may be a suitable alternative.

Monitoring system, a static camera is working to collect surveillance video wall mounted. Commonly used to obtain background subtraction is adopted human moving around, and the background pixel is estimated by non-parametric method. The size of information and speed, altitude, aspect ratio and speed information, including newly defined feature, CCMEI are used to indicate activity. In the recognition phase, the activity classes SVM decision tree is used to learn the limits. All classes are different from the one considered in the previous approach (Wang et al, 2007; .. et al, 2005),[8] we have a hierarchical manner activity classification. , an SVM binary decision tree classifier at each node, and the SVMs binary tree architecture (SVM-BTA). A multi-class support vector machine classifier with designated are aggregated to form a binary tree. The proposed monitoring system label only activity, but also detecting unusual activity decline and implementing intelligent home care for the asymmetry cannot issue a warning.

Proactive computing situations such as health care or life care requires active supporter of the anticipated public and takes appropriate action on their behalf, that is the technology applied and many researchers, institutions and commercial companies draw attention.

1.2 Problem Statement

This dissertation focuses on the problem of action recognition in realistic video material, such as movies, Internet and surveillance videos. In order to be more precise about our goal, we clarify the meaning of action and activity recognition.

Human action is composed of sequences, which are themselves structured with poses. In order to detect action recognition human blob detection is primary requirement. The action video always contains shape descriptor that can be used for detection of the poses. By detecting poses and putting them in order provides a sequence. This sequence is used for identifying the action. Pose detection and their sequence two are different in sense that if we maintain order the temporal component

does not gets distorted.

In this sense, an action can be precisely localized in a short interval in time, yet it can also refer to an event that lasts for a rather long time period. For clarification, action taxonomy can be defined as action primitive (or movement), action, and activity. An action primitive describes a basic and the atomic motion entity out of which actions are built. An activity is a set of several actions. Activities can be understood as larger scale events that often depend on the context and the environment in which the action happens.

1.3 Aim

This project is about Activity recognition and all the analysis and application is done over KTH dataset as described above. The six actions are recognizing though different algorithms. We have used Support Vector Machine to classify different actions of the dataset. We used Nearest Neighbor Algorithm also in individual and with the SVM. Before it different feature dimension reduction techniques are used as Linear Discriminant Analysis and Principal Component Analysis. Their results are compared to each other and to other methods of similar kind and to the same data set. KTH human motion dataset of six actions (walking, jogging, running, boxing, hand waving and hand clapping) performed by 25 different artists, are included. Used four different scenarios: outdoors, outdoors with different fabrics with zooming outdoors and indoors. Vary considerably in performance and duration, and to some extent in the approach. Background are relatively stable. In addition to zooming, landscape, only slight camera movement.

1.4 Context

Numerous works and methods have been proposed in the past within the field of action and activity recognition. Since recognizing actions in videos are a challenging problem, a lot of approaches have considered simplified settings. For a broader view, we discuss existing works according to the type of video data that they employ. For this, we distinguish the categories “controlled video data”, “constrained video data”, “uncontrolled video data”.

Controlled video data Controlled video data is acquired in a way to facilitate its automated processing. For instance, markers can be attached to human actors for detecting joints and limbs. Lighting conditions can be controlled to better detect markers and human bodies; multiple cameras can be placed in order to cover a necessary range of viewpoints for 3D reconstruction. A prominent example is commercially high-end motion capture systems for film productions. These use extensively optical markers and a large set of cameras to record motion up to the level of facial gestures and finger movements.

Constrained video data. Applications that operate on constrained video data are able to influence environmental parameters to a limited degree. This is the case for commercial video game platforms based on visual interfaces, such as the Project Natal [Microsoft, 2009][9]; certain assumptions can be made, e.g., a single person fully visible or favorable lighting conditions. However, certain robustness is necessary with respect to other visual conditions (e.g., varying size of humans, different clothing, motion variability) that cannot be influenced.

Uncontrolled video data. Uncontrolled video data is recorded under conditions which cannot be influenced. This is the case for, e.g., TV and cinema style movie data, sports broadcasts, music videos, or personal amateur clips. Only very few assumptions, if any, of a rather general nature can be made, such as humans are present and relative well visible. The main challenges for this more realistic data include changes of viewpoint, scale, and lighting conditions, partial occlusion of humans and objects, cluttered backgrounds, abrupt movement etc.

1.5 Outline of Thesis

The thesis is organized in the following way:

1. Chapter 1: Introduction introduces the context and motivation of the research presented in this thesis.

2. Chapter 2: Related work describes related work and our contribution.

3. Chapter 3: Methodology explains about different approaches used in detail. It tells about dimension reduction methodologies like PCA and LDA.

4. Chapter 4: This chapter describes the classification methods for activity recognition. Since SVM and KNN only are used here these only are described.

4. Chapter 5: Experiments and Results describe the datasets used for HAR, briefly present our algorithm and in the end we describe the parameter values and scenarios in which experiments were performed.

5. Chapter 6: Conclusion and Future Work concludes the project's thesis along with future plans.

CHAPTER 2

LITERATURE REVIEW

Here, silhouette-based recognition we have used in our method. Used silhouette is the basic unit of recognition. From the beginning of the activity recognition is used as the basic element. Also developed new methods in the later stages. Here is a discussion of related work is related to our method and our method have to remember that all these demands of research. SEI silhouette energy image for image recognition described in[10] by Ahmend et al, and variable action model is developed to detect human action. And claims are for non-stationary camera. Tseng et al. [11] Spatio- temporal information used for action recognition. Tseng et al. A three-step process k-NN classification purposes and for temporary data spatio- an efficient differential is used for the construction of temporary subspace in which [12] proposed silhouette-based human action recognition system to fight the difference between a non-base vector (extracted using are NCDAV) method. Rahman et al. [13] To approach the idea of negative space is proposed a method for dealing with change. It is also one of the major challenges to human action recognition that solve the problem of long shadows. Charooui a cooperative evolutionary algorithm based on the optimization of a human action recognition method is proposed, where a method returns. Coevolution through three different populations, for example, in relation to convenience and selection parameters is to achieve the best-performing individuals are developed. The fitness function is based on the result of human action recognition method. Riemannian geometry of shape spaces used by Abdelker et al. [14] A temporal sequence of human silhouette to represent a human gesture. Red et al. [15] Used to decrease the size and dimensions based on PCA features. He is lying markerless video analytic appearance-based features, enhances the movement of the body parts have been removed from the system is proposed.

2.1 Different Approaches for Activity Recognition

To this end, we constructed structure existing in three categories:

- **Human model based methods (Section 2.1.1)** of human body parts in a full 3D (or 2D) model employs recognition and action movements with the body part position information is used.
- **Holistic way (Section 2.1.2)** knowledge about localization in video using human body parts and consequently without any notion of the attribute, indicating that the global body movements to learn an action model.
- **Local feature methods (section 2.1.3)** are entirely based on descriptors of local regions, no prior knowledge about human positioning nor of any of its limbs is given.

General action and activity recognition and body tracking motion analysis as well as on the survey include Weinland et al. [16], Poppe [17], Moeslund et al. [18], Buxton [19], Moeslund and Granum [20], Gavrila [21], Aggarwal and Cai [22]. In addition, Hu et al. [23] Offer a survey for video surveillance, and Turaga et al. [24] For the analysis of high-level state-of-the-art review activity. Most relevant in our context are surveyed by Weinland et al. [16] And Poppe [17] which focused on the recognition of actions and action primitives.

2.1.1 Human model based methods

Model-based methods such as the human body positions and movements as part of the tasks identified employment information. A significant amount of research Moeslund et al., [18] As a prior model, with or without human kinematic joint positions on the human body, body parts, or recognition operations using the trajectories of landmark points, dedicated to If, Ali et al., [25] that can (see figure 2.1). Localization of body parts in films (Ferrari et al. [25], Eg, Ramanan et al.[13]) Has been investigated in the past and some works have shown impressive results.

2.1.2 Holistic methods

The overall method does not require the localization of body parts. Instead, the structure and dynamics of the global body is used to represent human actions. Polana and Nelson [27] "without finding his body parts getting their man" referred to this approach. Frankly about body parts or the information using a kinematic model approach than the overall global motion and appearance information model

representations, since they are very simple. So their calculations are more efficient as well as robust general. This aspect of the background clutter, camera ego motion, and occlusion of body parts localization is particularly difficult to render realistic is particularly important for video.

In general, the overall approach can be broadly divided into two categories. To represent first-class functions, background subtraction masks or shape or silhouette information stemming from the difference images are employed. The second category is based primarily on size and optical flow information.

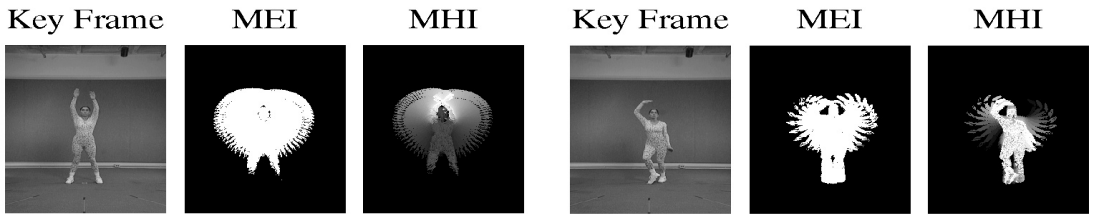

Fig. 2.1: Shape masks from difference images for computing motion history images (MHI) and motion energy images (MEI) Yilmaz [1]

Shape mask and silhouette based methods

Several approaches for action recognition to represent the dynamics of the human body and the human shape your silhouette masks and use information. Yamato et al. [26] silhouette images (cf. figure 2.1) are among the first to propose. Silhouette their representation on a grid for each cell count and calculate the ratio of foreground to background pixels. Grid representations are quantized in a vocabulary, and tennis action then hidden Markov model (HMM) [28] use the "word" as the visuals are learning.

Bobick and Davis [29] difference images to detect human actions shape the use of masks. As illustrated in figure 2.1 as represented in the action, the authors, the so-called motion energy images (MEI) and motion history images (MHI) at work. More precisely, MEIS indicate that binary masks areas of speed, and they (the more recent, more weight) according to the point in time when these areas MHIs weight. This approach for action recognition is the first to introduce the idea of temporary templates.

Tennis player silhouette of the information office of the joints between the main frames are tracked. This approach, as such, can be applied to 3D animation that allows inferring the position of body parts.

Silhouette information shapes the space-time-action model is introduced by Blank et al. [30], Gorelick et al. [11]. Silhouette information is by using background subtraction. Figure 2.2 shows some examples of space-time shapes. Writers such local saliency, action dynamics, shape structure and orientation of the Poisson equation to extract features such as the use of the properties of the solution. A high dimensional feature vector that describes the amount of 10 frames in length, during classification, these quantum space-time images of test sequences are matched in a sliding window fashion.

Man uses the size of the space-time that a job Yilmaz and Shah [1] is proposed. Space to clear temporary shape similar background subtraction et al, are derived from the contour information. For a strong representation, the action on the surface of the shape (eg saddle, valley, ridge, peak, pit points) is represented by sets of feature points. To identify tasks, the authors point-to-point correspondence temporary space match the size of the proposed holography

Figure 2.2: Space-time volumes for action recognition based on silhouette information (courtesy of Blank et al. [2005]).

Weinland and Boyer [16] introduce an orderless representation for action recognition using a set of silhouette exemplars. Action sequences are represented as vectors of minimum distance between silhouettes in the set of exemplars and in the sequence. Final classification is done using Bayes classifier with Gaussians to model action classes. In addition to silhouette information, the authors also employ the

Chamfer distance measure to match silhouette exemplars directly to edge information in test sequences.

Optical flow and shape based methods

Human-centric approaches based on optical flow and generic shape information form an- other sub-class of holistic methods. As one of the first works in this direction, Polana and Nelson [32] propose a human tracking framework along with an action representation using spatio-temporal grids of optical flow magnitudes as shown in figure 2.2. The action descriptor is computed for periodic motion patterns. By matching against reference motion templates of known periodic actions (e.g., walking, running, swimming, skiing) the final action can be determined.

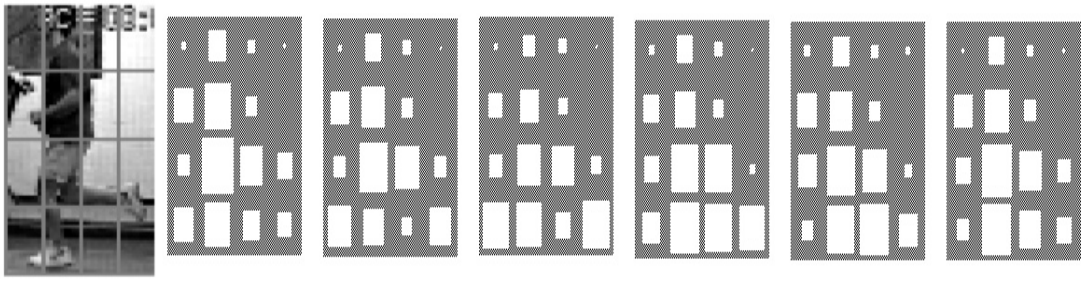

Figure 2.3: A human centric grid of optical flow magnitudes to describe actions (courtesy of Polana and Nelson [32]).

In another approach purely based on optical flow, Efros et al. [33] track soccer players in videos and compute a descriptor on the stabilized tracks using blurred optical flow.

Their descriptor separates x and y flow as well as positive and negative components into four different channels, as can be seen in figure 2.7. For classification, a test sequence is frame-wise aligned to a database of stored, annotated actions. Further experiments include tennis and ballet sequences as well as synthesis experiments.

The same human-centric representation based on optical flow and human tracking for action recognition, is employed by Fathi and Mori [31]. As classification framework, the authors use a two-layered AdaBoost, Freund and Schapire, [34]

variant. In a first step, selecting discriminative pixel flow values in small spatio-temporal blocks learns intermediate features. The final classifier is then learned from a all previously aggregated intermediate features. Evaluations are carried out on four datasets: KTH, Weizmann, a soccer, and a ballet dataset.

Rodriguez et al. [35] propose an approach using flow features in a template-matching framework. Spatio-temporal regularity flow information is used as feature type. Regularity flow shows improvement over optical flow since it globally minimizes the overall sum of gradients in the sequence. Rodriguez et al. learn cuboid templates by aligning training samples via correlation. For classification, test sequences are correlated with the learned template via generalized Fourier transform that allows for vectorial values. Results are demonstrated on the KTH dataset, for facial expressions, as well as on custom movie and sports actions.

To localize humans performing actions such as sit down, stand up, grab cup and close lap- top, Ke et al. [36] use a forward features selection framework and learn a classifier based on optical flow features.

A method purely based on shape description is presented in Lu and Little, [37]. In their experiments, Lu and Little track soccer or base-ball players and represent every frame by a descriptor using histograms of oriented gradients. They then employ principal component analysis (PCA) to reduce dimensionality. An HMM is used with a few states models actions such as running/skating left, right etc.

Hybrid representations combine optical flow with appearance information. Schindler and van Gool [38] use optical flow information. Majority voting yields a final class label for a full sequence in multi-class experiments. Results are carried out on the KTH and Weizmann dataset.

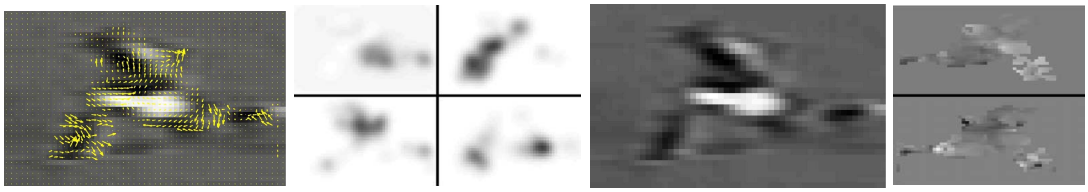

Figure 2.4: Motion descriptor using optical flow: (a) Original image, (b) Optical flow, (c) Separating the x and y components of optical flow vectors, (d) Half-wave rectification and smoothing of each component.

2.1.3 Local feature methods

For a local area of spatio-temporal features video contains description shape and motion information. They view their spatio- with respect to temporary shifts and scales of events with multiple offers background clutter and provide a relatively independent representation. In the following, we first discuss existing spatio-temporal feature detectors and feature descriptors. Methods based on feature trajectories are presented separately since their conception differs from space-time point detectors. We then review approaches which employ the orderless bag-of-features representation and which build spatio-temporal action models based on local features. Finally, methods for localizing actions in videos are discussed.

Feature detectors

Feature saliency detectors specific actions by maximizing specializing in video usually temporary spatio- select locations and scales. Laptev and Lindeberg [41] Laptev corner ness Harris criterion Harris and Stephens, [32] based on a spatio-cosmic expansion are the first to propose a feature detector. At one point criterion Corner ness, spatio-temporal each video is based on temporal eigenvalues of the second moment matrix. Local Maxima indicate points of interest.

Using entropy of a space-time extension of a major area detector, introduced by Los et al. [39]. Entropy of a video sequence for the temporal derivative of the space-time around the situation is in a cylindrical neighborhood. A sparse representation to obtain more stable interest points, local Maxima thresholded candidates and groups,.

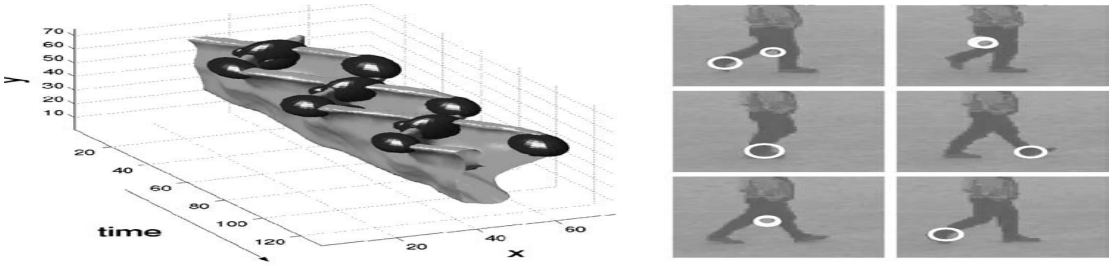

Figure 2.5: Spatio-temporal interest points; (left) 3D plot of a leg pattern (upside down) and the detected local interest points; (right) interest points overlaid on single frames in the original sequence

Most feature detectors determine the saliency of a point with respect to its local neighborhood. Wong and Cipolla [40] suggest determining salient features by considering global information. For this, video sequences are represented as dynamic texture with a latent representation and a dynamic generation model. This allows to synthesize motion, but also to identify important regions in motion. The dynamic model is approximated as linear transformation. A sub-space representation is computed via non-negative matrix factorization. Local 2D interest in the sub-space images and temporal maxima in their coefficient matrix indicate localizations of globally salient positions

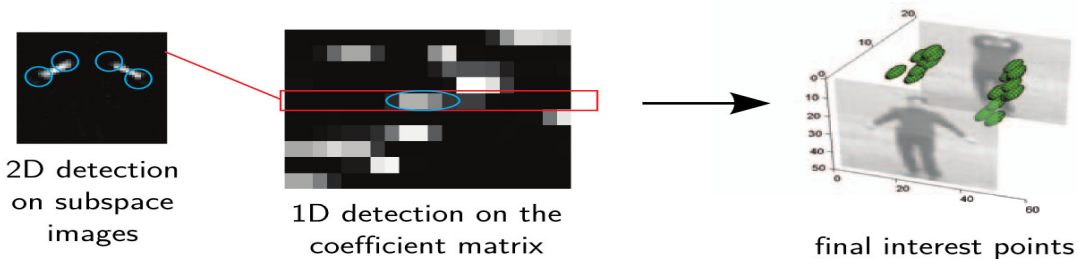

Fig. 2.6: Feature detection with global information; (left) spatial feature positions are given by 2D detections in subspace images, (middle) the temporal position is given by maxima in the coefficient matrix; (right) final positions in a waving sequence.

Feature descriptors

Feature descriptors capture shape and motion information in a local neighborhood surrounding interest points. Among the first works on local descriptors for videos, Laptev and Lindeberg [41] develop and compare different descriptor types: single- and multi-scale higher-order derivatives (local jets), histograms of

optical flow, and histograms of spatio-temporal gradients. Histograms for optical flow and gradient components are computed for each cell of a M × M × M grid layout describing the local neighborhood of an interest point. A different variant describes the surrounding of a given position by applying PCA to concatenated optical flow or gradient components of each pixel. The resulting descriptor uses the dimensions with the most significant eigenvalues. In their experiments, Laptev and Lindeberg report best results for descriptors based on histograms of optical flow and spatio-temporal gradients.

In a similar work, Doll´ar et al. [42] evaluate different local space-time descriptors based on brightness, gradient, and optical flow information. They investigate different descriptor variants: simple concatenation of pixel values, a grid of local histograms, and a single global histogram. Finally, PCA reduces the dimensionality of each descriptor variant. Overall, concatenated gradient information yields best performance.

Bag of features

Bag features a popular representation based on local features (both) model. The haphazard methods are a popular choice for representing textual data retrieval applications where the document originates. The frequency distribution of words in terms of the bag model describes the domain as text documents [Salton, 1968] has been implemented on a large scale.

Visual recognition tasks, Cula and Dana [23], Sivic and Zisserman [43], Sivic et al. [34] respectively texture classification, object / scene retrieval, image classification and object localization, classification for applications with a view to extend this concept to one of the first authors. Schu LDT et al. [32], dollar et al. [42], Niebles et al., [24] the first extension proposed for action recognition.

In the video for the representation, the feature detectors to determine a set of leading positions in the scenery. The Feature position descriptors for the local neighborhood are to calculate a vector representation. Visual vocabulary (or codebook) is calculated by applying a clustering algorithm (eg, K-means) feature descriptors derived from sequences on training; each cluster is referred to as the visual

word. Descriptors are quantized by assignment to the closest visual word, and visual terms the phenomenon of video sequences are represented as a histogram.

Spatio-temporal action models

Since the BoF model does not incorporate any geometrical information between features, recent works propose methods to build stronger action models based on local features. For instance, Laptev et al. [41] include weak geometric information by introducing rough spatio-temporal grids overlaid on video sequences. Grid layouts as well as shape and motion descriptors are combined by kernel fusion using a non-linear SVM. A greedy optimization strategy learns the best combination of grids and feature types per action class. The authors demonstrate the effectiveness of their approach on the KTH dataset and a large set of sample actions obtained from Hollywood movies.

Han et al. [44] combine different local features with varying layouts and types (histograms of oriented gradients, histograms of optical flow, histograms of oriented spatio- temporal gradients) by fusing multiple kernels using Gaussian processes. By employing various object detectors (for full person, upper body, chairs, cars), they additionally include information about the absence or presence of objects in the sequences. Results on different datasets (KTH, Hollywood1, Hollywood2) demonstrate state-of-the-art classification results.

Action localization by voting

Combined with a voting scheme, local features can also be employed to spatially as well as temporally localize actions in videos. For instance, Niebles et al. [45] perform a latent topic discovery and model the posterior probability of each quantized feature for a given action class. In order to localize actions, features are spatially clustered in each frame using k-means.

In order to localize actions in YouTube video sequences, Liu et al. [37] propose an approach based on pruning local features. First, spatio-temporal features are detected and their mean position over a range of neighboring frames is computed. Features that are too far away from the center position are pruned. Second, static

features are computed over all frames. By applying the PageRank algorithm over a graph for feature matches in a video sequence, the authors are able to identify discriminative features. For this, similar background features are assumed to by less frequently visible than foreground features. Finally, static and motion features are combined with an AdaBoost classifier. Action localization is carried out with a temporal sliding window over spatio-temporal candidate regions defined by the center and the second moments of motion as well as static features.

CHAPTER 3

PROPOSED METHODLOGY

The proposed method is explained with the help of flow diagram shown in fig. 3.1. The first step is for extraction of frames from the video. The frames, which are extracted here, are the frames that have content that contains maximum size of silhouette. The first 31 frames that have maximum are content are selected. And further silhouette and feature extraction techniques are used. And then dimension reduction and classification is used. This all is described in following steps.

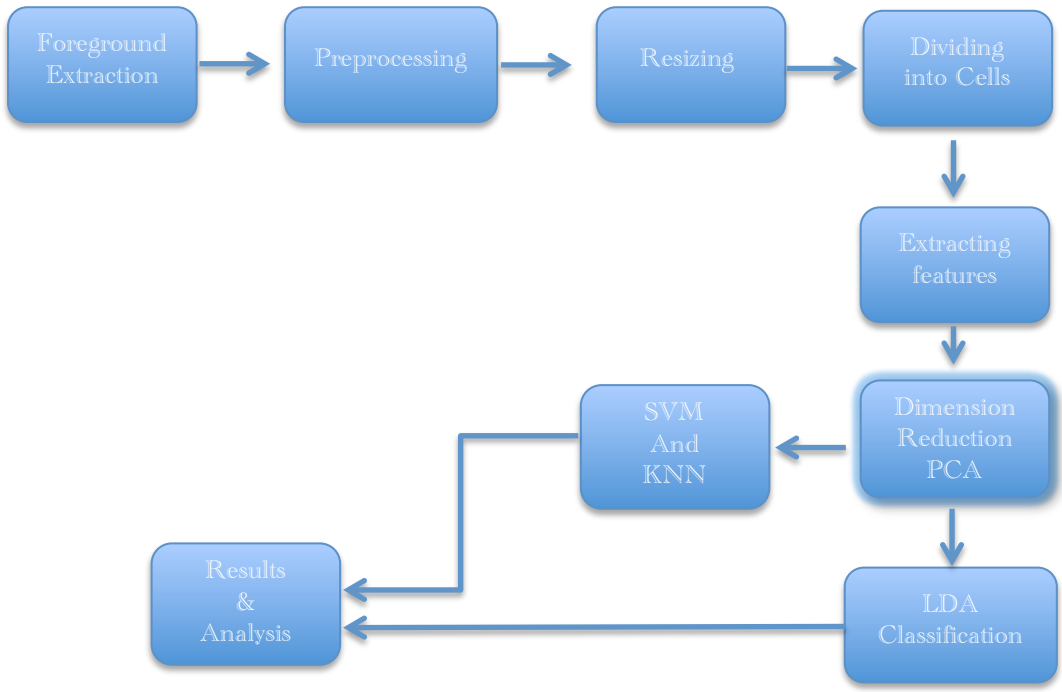

Fig 3.1: flow diagram of algorithm used

3.1 Feature extraction:

Based on recent developments of visual recognition in static images, many concepts have been successfully extended to video sequences, for instance: feature detectors, feature descriptors, bag-of-features representations, local features based

voting for localization. In this chapter we discuss about feature representation and feature extraction. Before we move further should discuss some basic steps.

3.1.1 Accessing the video and Extracting Key Frames

Feature extraction is the first requirement for analysis of data. All the analysis is done over feature set collected and based on that one can find result by applying different techniques. Following dig. Shows how frames and significant images have been selected out of large number of frames inside a video.

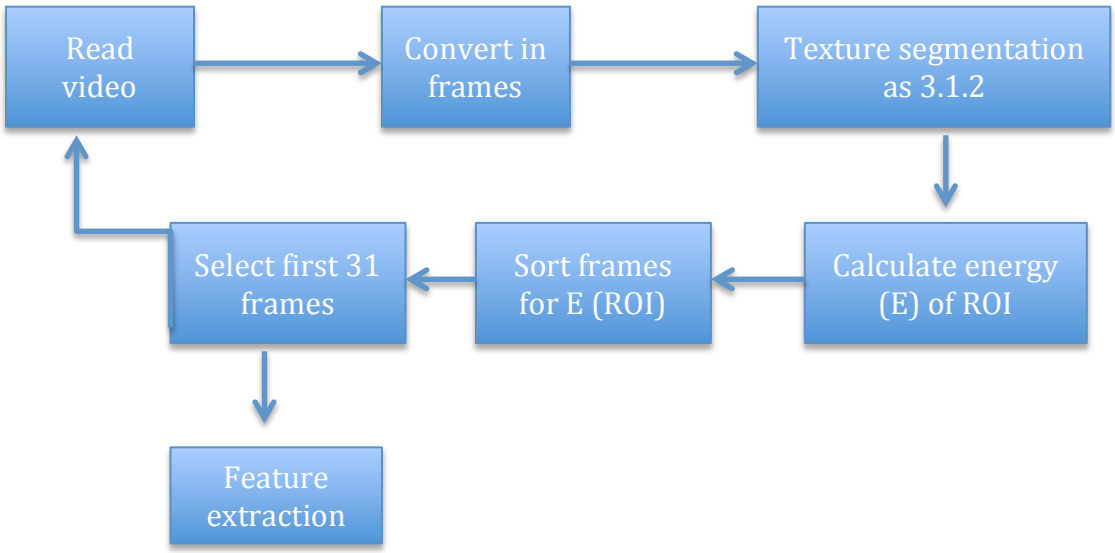

Fig. 3.2 Method used for significant frame selection

The frames, which have largest energy are selected for further processing. Energy is calculated as the no of pixels that are non-zero . This is basically size of objects found inside the frame

3.1.2 Texture based segmentation for silhouette extraction:

Texture based segmentation for silhouette extraction is done is over famous dataset of KTH that is having different situations and different scenarios. Some of little preprocessing is done to make silhouette with minimum noise though dataset, being shot in different lightning conditions full accuracy is not achieved. This inaccuracy and noise becomes limitation for classification and hence two classifications are implemented to minimize effect of noise in further stages.

Texture filter (Gray-Level Co-Occurrence Matrix Descriptor): A statistical method of examining texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. Gray level co-occurrence matrices (GLCM) is a method which allow to describe texture based on its differences in intensity towards different directions. It was proposed in [46] by Robert M. Haralick, that is why it is also called Haralick features. He proposed to count pixel pairs gray level occurrences in analyzed image. Such statistics, gathered for specific locations on rectangular grid, allows extracting many interesting characteristics**.**

The idea is presented in Figure 3.3. At the first step, the gray-scale of original image (Figure 3.3 a) is reduced (Figure 3.3 b) to N levels. Then for each pixel (denoted as i) from the region of descriptor computation a corresponding pixel (marked as j) is appointed lying on the direction given by an angle α and at the distance of d (Figure 3.3 c). Most commonly used are: distance one pixel and angles of 0, 45, 90 or 135 degrees, covering 3x3 neighborhood. The values of each pixels pairs i, j and j, i are used as an indexes to co-occurrence matrix (M(i,j)) locations where information about co-occurrences are accumulated (Figure 3.3 d). After building a co-occurrence

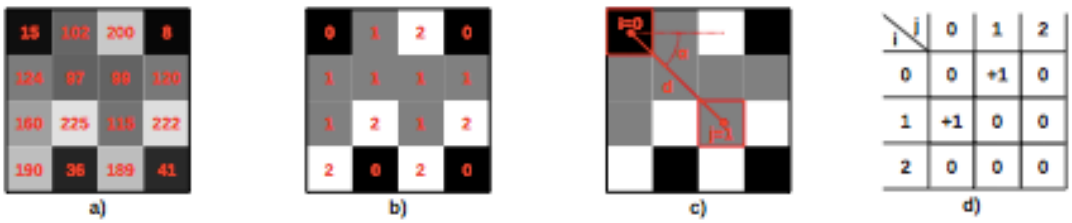

Fig. 3.3: Co-occurrence matrix computation, a) original image, b) reduced grayscale image, c) single pixel co-occurrence pair determining, d) co-occurrence matrix incrementation

Matrix for interesting area, the probability matrix is computed based on Equation 3.1:

$$P(i,j) = \frac{M(i,j)}{\sum_{i,j} M(i,j)} \tag{3.1}$$

Simply by dividing each value of the co-occurrence matrix by the sum of all its elements.

Based on the probability matrix computed from the co- occurrence matrix, different features are computed. The full list of features is presented in [46], some of them are listed below:

Energy:

$$f_1 = \sqrt{\sum_i \sum_j P(i,j)^2} \quad (3.2)$$

Homogeneity:

$$f_2 = \frac{P(i,j)}{1+(i-j)^2} \quad (3.3)$$

Entropy:

$$f_3 = -\sum_i \sum_j P(i,j)[\ln P(i,j)] \quad (3.4)$$

Contrast:

$$f_{4=} \sum_i \sum_j (i-j)^2 P(i,j) \quad (3.5)$$

3.1.3 Masking image with Texture filter

This kind of architecture defined for any type of mask or filtering and cross-correlation for feature detection can be a highly parallel implementation. Much of the architecture and location of psychological edge mask is made between zero crossing Fig 3.3(3) is made in the notes as if it were pulled apart to exaggerate the intensity difference between the two segments of humans an edge between two regions as if it were pulled apart to exaggerate the intensity difference as is done in Fig 3.3 (3). Rough and smooth texture difference provide us the result.

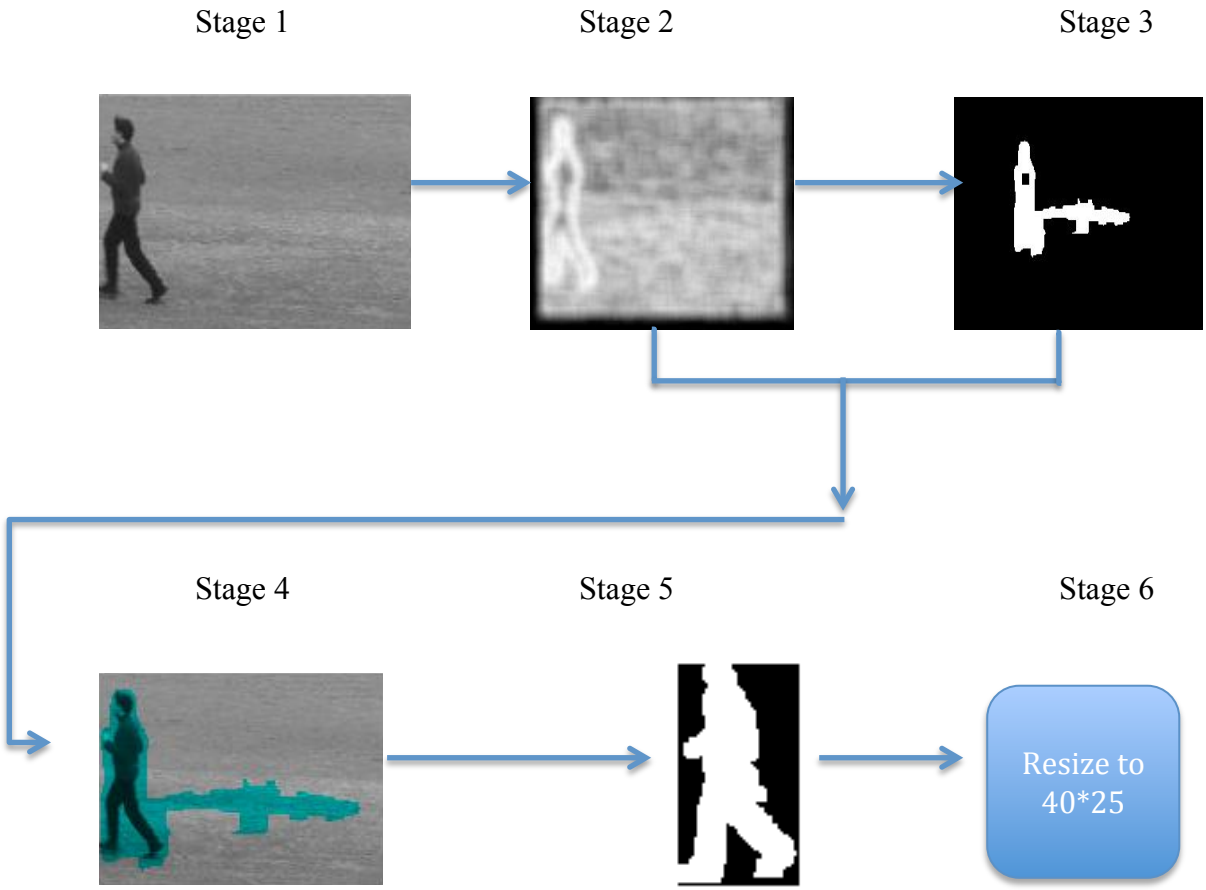

Fig 3.3 Extraction of silhouette stage (1). Original image (2). Texture mask (3), (4). Images selected by rough mask (5). Final silhouette extracted from original image (1)

3.1.4 Selecting the image with largest bounding-box

Since there is definitely a confusion that which part is human silhouette and which part is not. To select this we go for a bounding box area. And a part of image is selected if it has largest part and labeled it as human silhouette. As in image (4) we have two parts shown as same texture but human part has large are and therefore it has been selected as silhouette.

3.1.5 Resizing Silhouette Image

Silhouettes of actors extracted from earlier steps are not in uniform size and that makes necessary to resize the image. After resize size of becomes 40×25 and this always maintain a ratio of 8:5. This ratio is enough to fit information contained in a silhouette whether it is of any action e.g. jogging, jumping etc. that are in the KTH

dataset. Interpolation is the method used to estimate pixel at a location in between other two or many pixels. Whenever image is enlarged, the output image contains more number of pixels than the previous one from which new one is derived. And the interpolation is used to determine the values for the additional pixels. In general, every pattern recognition system, a preprocessing step is necessary as the temple. We have a certain width x height size of the bounding rectangle is to normalize the binary silhouettes. Movements of the hands and feet are self-occluding motion to express the human body, silhouette normalization is a difficult task. We stand for the silhouettes dependent normalization technique is needed.

Image interpolation between pixels in a position is to estimate the value of a process image. Image detail is an image; the output contains more pixels and hence the data than the input image. For additional pixels interpolation is used to determine values. Weightings of each pixel are based on distance from the point. To find the values of the image element in output image uses image 'bicubic' interpolation size, but the other struck a bilinear interpolation method to be used can use interpolation methods.

If you reduce the size of an image, because there is less pixels in the output image, you may lose image content that is some of the pixels and this phenomenon can cause aliasing. As a result of the reduction in size that would normally produce aliasing in the image "stair step" pattern, or as ripple effect pattern appears also called as Moire effect.

3.1.6 Dividing into cells: After resizing we got image of 40×25. That contains total pixels equal to 1000. We divide this picture into grids of 5×5 image. Since we have already converted image into binary form, we can calculate number of white pixels in this cell and this number is used as feature for this particular cell or grid. Total 8×5=40 number of features comes out from this image.

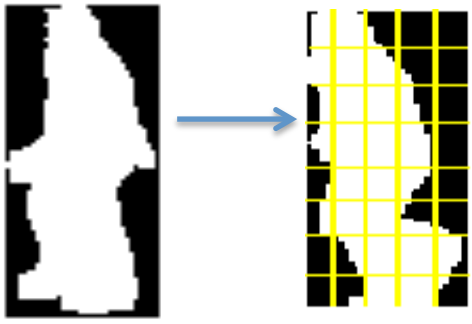

Fig 3.4 Dividing image into parts called cells

3.1.7 Feature Extraction from Grid Image

Total 8×5=40 number of features comes out from this image. One image generates equal to 40 features. Now it depends on the number of frames that are key to the action or the activity. We have used just 31 frames out of a video that represents an activity. Hence the total features generated from an action video are 31×40=1240. As described above all these features contain the number of white pixels according to particular cell or grid. Division of image is shown in fig 3.4. It is shown that image has 40 cells and f1 is number of white pixels in this cell.

Features generated from an image

$$f_{i1} = \{\text{f1, f2 f3 } \ldots \ldots \text{f40}\}_{i=1}$$

Similarly we have

$$f_{i2} = \{\text{f1, f2 f3 } \ldots \ldots \text{f40}\}_{i=2}$$

:

:

:

$$f_{in} = \{\text{f1, f2 f3 } \ldots \ldots \text{f40}\}_{i=n}$$

Here n = number of frames in particular video representing activity.

Final feature set from a video

$$F_{jo} = \{f_{i1}, f_{i2} \dots \dots \dots f_{in}\}_{a=jogging} \ ;$$

$$F_{hc} = \{f_{i1}, f_{i2} \dots \dots \dots f_{in}\}_{a=handclapping} \ ;$$

:

:

$$F_{ru} = \{f_{i1}, f_{i2} \dots \dots \dots f_{in}\}_{a=running} \ ;$$

White pixels=[20 22 10 0 0 ……..40 items] =f_{i1}

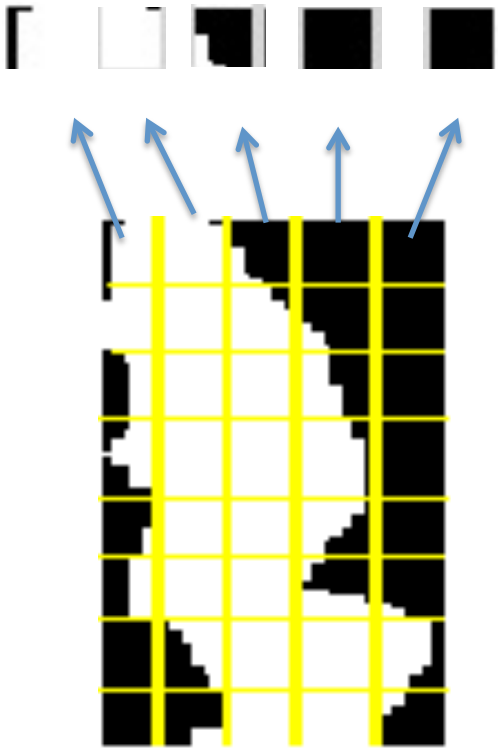

Fig 3.5 showing results from a running image

Here F_{jo} is for jogging and like wise we do have different feature vectors of dimension 1240. These feature vectors for all six activities are generated as shown in

this section. By combining all the six feature vectors of corresponding vector data matrix is generated. Since 80 videos have been used here for one activity. So in total such 80 multiplied by 6 total videos are generated and because each video used generates a feature vector, same no that is 480 features are generated.

3.2 Feature Representation

We have used just 31 frames out of a video that represents an activity. Hence the total features generated from an action video are 31×40=1240. As described above all these features contain the number of white pixels according to particular cell or grid.

Final feature set from a video

$$F_{jo} = \{f_{i1}, f_{i2} \dots \dots \dots f_{in}\}_{a=jogging} \ ;$$

$$F_{hc} = \{f_{i1}, f_{i2} \dots \dots \dots f_{in}\}_{a=handclapping} \ ;$$

:

:

$$F_{ru} = \{f_{i1}, f_{i2} \dots \dots \dots f_{in}\}_{a=running} \ ;$$

Here n is used for all activity video as 31.

After resizing we got image of 40×25. $That$ contains total pixels equal to 1000. We divide this picture into grids of 5×5 image. Since we have already converted image into binary form, we can calculate number of white pixels in this cell and this number is used as feature for this particular cell or grid. Total 8×5=40 number of features comes out from this image. One image generates equal to 40 features. Now it depends on the number of frames that are key to the action or the activity. We have used just 31 frames out of a video that represents an activity. Hence the total features generated from an action video are 31×40=1240. As described above all these features contain the number of white pixels according to particular cell

or grid.

Some of little preprocessing is done to make silhouette with minimum noise though dataset, being shot in different lightning conditions full accuracy is not achieved. This inaccuracy and noise becomes limitation for classification and hence two classifications are implemented to minimize effect of noise in further stages.

Since we have already converted image into binary form, we can calculate number of white pixels in this cell and this number is used as feature for this particular cell or grid. Total 8×5=40 number of features comes out from this image. One image generates equal to 40 features. Now it depends on the number of frames that are key to the action or the activity. We have used just 31 frames out of a video that represents an activity. Hence the total features generated from an action video are 31×40=1240. As described above all these features contain the number of white pixels according to particular cell or grid.

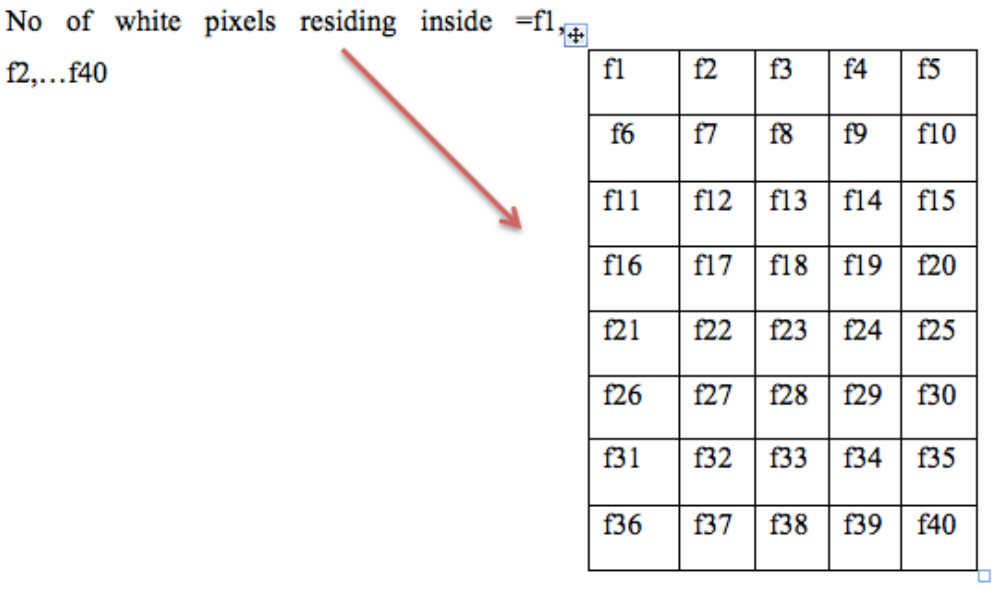

Fig 3.6 showing feature representation from an image

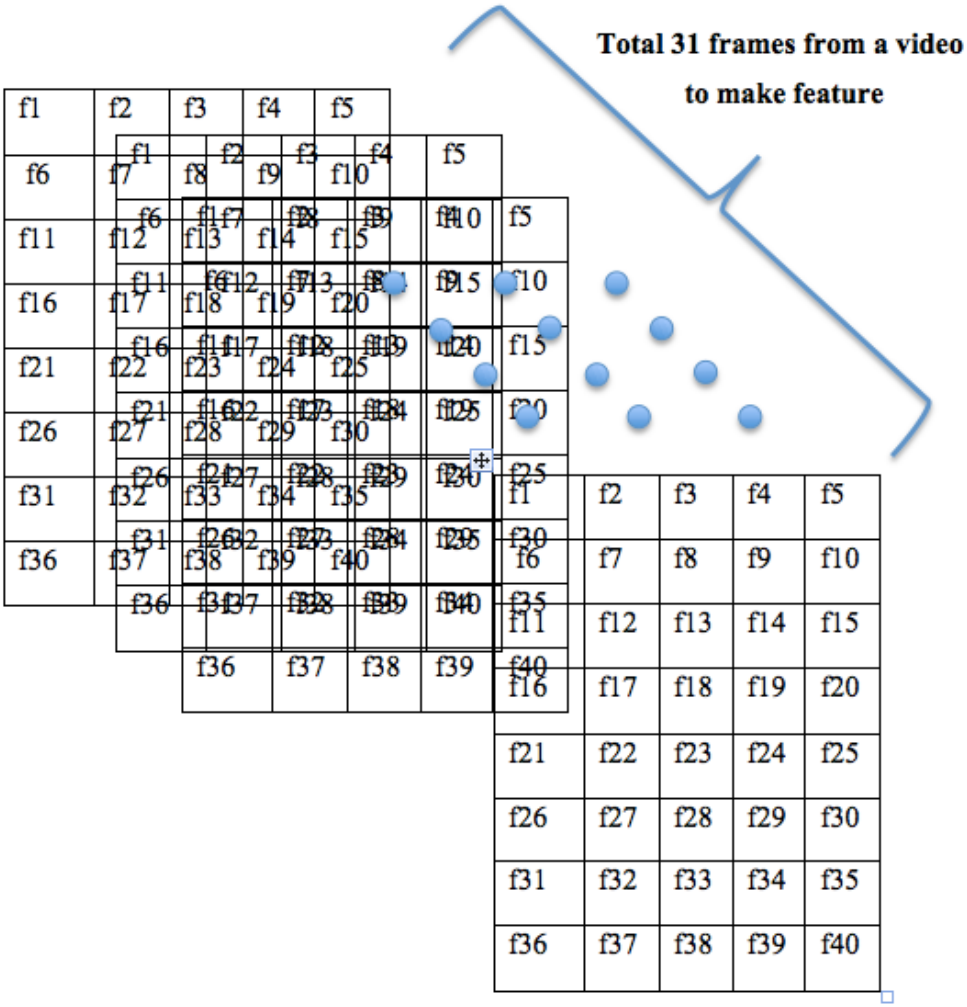

Fig 3.7 Collection of 31 key frames for generation of feature vector

As shown in all figures, Hence the total features generated from an action video are 31×40=1240. As described above all these features contain the number of white pixels according to particular cell or grid. Total 8×5=40 number of features comes out from this image. One image generates equal to 40 features. Now it depends on the number of frames that are key to the action or the activity. We have used just 31 frames out of a video that represents an activity. Hence the total features generated from an action video are 31×40=1240. As described above all these features contain the number of white pixels according to particular cell or grid. Division of image is shown in fig 3.3. It is shown that image has 40 cells and f1 is number of white pixels in this cell.

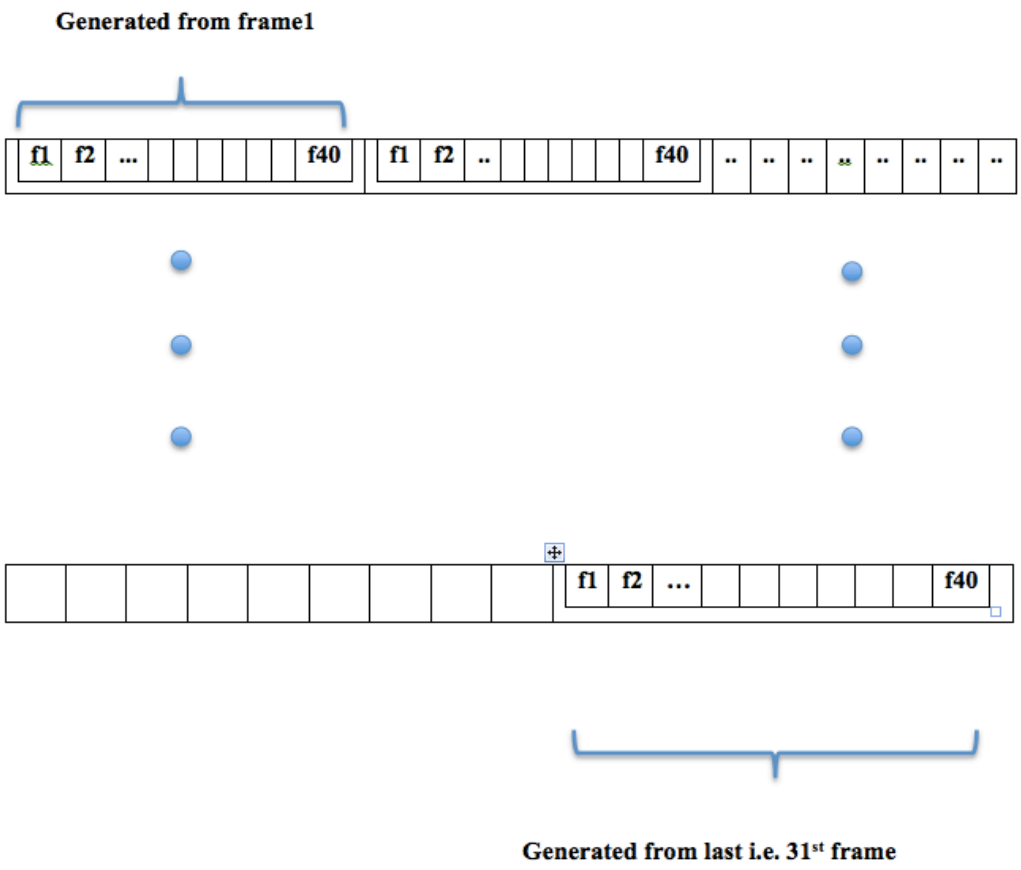

Fig 3.8 Final feature vector generated from an activity video

Now we do have different feature vectors of dimension 1240. These feature vectors for all six activities are generated as shown in this section. By combining all the six feature vectors of corresponding vector data matrix is generated. Since 80 videos have been used here for one activity. So in total such 80 multiplied by 6 total videos are generated and because each video used generates a feature vector, same no that is 480 features are generated. The data generated is of 480 rows and 1240 columns. This is a huge data set, for this to reduce to a lesser no of columns different dimension reduction techniques are used these are described in following steps.

3.3 Dimension reduction:

From previous steps we get no of features for an activity or action that is represented by some actor and thorough a video. This huge number of features makes the classification a time taking process. This can be reduced significantly if we have

less no of features. All of this can be done through various algorithms. Some linear dimension reduction techniques are as LDA (linear discriminant analysis), PCA (Principal Component Analysis) Turk et al. [47]. Some other algorithms are also there which are more efficient but are complex and the time taking. These technique are not being used as these defeat the purpose of fast and efficient too method that we have developed in this paper.

3.3.1 Principal Component Analysis

PCA is a variable's dimension reduction procedure. Principal component analysis of the variables used as predictor or criterion can be. In this case, redundancy, possibly because they measure the same construct some of the variables are correlated with each other that means.

Because of this redundancy, you may, in the principal components (artificial variables) observed in a small number of variables should be possible to lose that trust. It is a variable reduction method or technique; its exploratory factor analysis is matching in many ways. You have a large number of variables like here we have 1240 element in our feature vector, the data is obtained, and u find or think that there is some redundancy in those variables when it is useful. In fact, a principal component analysis conducted step when conducting an exploratory factor analysis are almost identical to those followed when pursuing. Because of this redundancy, Differences between these two processes after title are described in detail in the other parts of this section.

To define PCA we take a set of vectors of *p*-dimension of weights also called loadings $W_{(k)} = (w_1, w_2 \ldots\ldots w_p)_{(k)}$ that map all row vector X_i where i=1,…,p of **X** to a new vector of principal component *scores* $t_{(i)} = (t_1, t_2 \ldots\ldots t_p)_{(i)}$, given by $t_{k(i)} = X_{(i)}.W_{(k)}$. So that the individual variables of **t** considered over the data set successively inherit the maximum possible variance from **x**, with each loading vector w constrained to be a unit vector

Technically, a key component of improved weighted observed variables can be defined as a linear combination. To understand the meaning of this definition, this

is the first time a major component scores are calculated on the topic of how to describe that is necessary.

Principal component analysis (PCA) in the way of performance, given that a key component to calculate a score for each subject is possible. Let us have example to understand; in the previous case, each subject would score on only two components: a score that is on satisfaction with supervision component, and a score on satisfaction with salary component. Seven-questionnaire item concerning a given component of the actual score better weighted and then summed to calculate your score will be.

The first loading vector $\mathbf{w}_{(1)}$ thus has to be calculated as:

$$w_{(1)} = \arg\max_{\| w \| = 1} \left\{ \sum (t_1)_{(i)}{}^2 \right\} = \arg\max_{\| w \| = 1} \left\{ \sum (X_{(i)}.W)^2 \right\} \tag{3.6}$$

In matrix form eq (3.1) can be written as:

$$w_{(1)} = \arg\max_{\| w \| = 1} \left\{ \| X_w \|^2 \right\} = \arg\max_{\| w \| = 1} \left\{ w^T X^T X_w \right\} \tag{3.7}$$

Since $\mathbf{w}_{(1)}$ is a unit vector, it at the same movement also satisfies

$$w_{(1)} = \arg\max \left\{ \frac{w^T X^T X_w}{w^T w} \right\} \tag{3.8}$$

Rayleigh quotient is now here to be maximized. For a symmetric matrix, quotient's maximum possible value is the largest eigenvalue of the matrix, which corresponds to eigenvector **w.** This rule also apply over XTX . With loading $\mathbf{w}_{(1)}$ in our hands, the first component of a data vector $\mathbf{x}_{(i)}$ can be calculated as a score $t_{1(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}$ in the transformed co-ordinates, or as the corresponding vector in the original variables, $\{\mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}\} \mathbf{w}_{(1)}$.

Rest all components can be found as:

The k^{th} component of vector can be calculated by subtracting the first $k - 1$ principal components from **X**:

$$\widehat{\boldsymbol{X}}_{\boldsymbol{k-1}} = \boldsymbol{X} - \sum_{\boldsymbol{s=1}}^{\boldsymbol{k-1}} \boldsymbol{X}\boldsymbol{w}_{(\boldsymbol{s})}\boldsymbol{w}_{(\boldsymbol{s})}^{\boldsymbol{T}} \quad (3.9)$$

and then calculating the loading or weight vector which provides the maximum variance in this new data matrix

$$W_k = \arg\max \{\| \widehat{X}_{k-1} w \|^2\} = \arg\max\{\frac{w^T \hat{X}_{K-1}^T \hat{X}_{k-1} w}{w^T w}\} \quad (3.10)$$
$$\| w \| = 1$$

By this we can find the rest all eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, with the maximum values corresponding to their eigenvalues and eigen vector. The k^{th} principal component of a data vector $\mathbf{x}_{(i)}$ will be given as a score $t_{k(i)} = \mathbf{x}_{(i)}.\mathbf{w}_{(k)}$ in the transformed co-ordinates, or as the corresponding vector in the space of the original variables, $\{\mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)}\}\ \mathbf{w}_{(k)}$, where $\mathbf{w}_{(k)}$ is the k^{th} eigenvector of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. The full principal components decomposition of $\mathbf{X}$ can therefore be given as $\mathbf{T}=\mathbf{XW}$ where $\mathbf{W}$ is a p-by-p matrix whose columns are the eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$

An introductory treatment of the subject, this will provide a comprehensive discussion. Instead, specific recommendations are often consistent with practices followed in Applied Research, will be created.

Step 1: The number of variables that are to be analyzed, initial removal of the components of the principal component analysis, may be equal to the number of components extracted. Six variables (that is our six classes of action) are analyzed in the present case, the six components may be removed. The first component of PCA to account for a significantly greater amount evolved in total variance can be expected. Successful progressively smaller amounts of each component will account for variance

Step2: We need to decide just how to feel worthy of rotation and retained for interpretation. In general, only the first few components will account for a significant amount of variance, and will account for the variance components in obscurity there

is hope. The next step of the analysis, therefore, must be retained for interpretation is to determine how many meaningful components. In this section it can be used in making decisions that will describe four criteria these are as follows

A eigenvalue one criterion. In PCA the number of components to solve the problem one of the most commonly used criteria Kaiser criterion [48], known as a criterion eigenvalue. With this approach, you maintain and greater than 1.00 any component with an eigenvalue interpretation. Total available variance in the observed dataset for each variable contributes one unit of variance. And eigenvalue greater than 1.00 that any component was contributed by a variable than is accounting for a large amount of variance. Such a component of the variance accounting for a significant amount, and is worthy of being maintained.

B Scree test. Scree Test described by [49], with the plot of eigenvalues, which are associated with every component and component with relatively larger eigenvalues and among those with smaller eigenvalues, a "break" to look for. The break is considered meaningful and are kept for rotation components appear before; those eigenvalues which appearing after the gap are assumed to be insignificant or not maintained.

The term "macadam" lies at the base of a cliff that shows loose debris. On top of a brake (cliff edge) after some meaningful components eigenvalues, would be: When a scree test, you will normally take the form of a rock scree plot is expected. Scree will lie rock bottom: trivial components eigenvalues.

Step 3: A final solution to the rotation Factor loading patterns and factor. After removing the initial components un-rotated factor pattern matrix will create one aspect Proc. There are entries in the factor-loading matrix. A factor loading of a factor pattern matrix (fpm) or structure matrix appears in a factor that is a general term for a coefficient. Oblique (correlated) components results in an analysis, the definition of a factor loading or factor in a factor pattern matrix based on the matrix structure is different. However, the situation (as the current chapter) orthogonal components results in an analysis that is simple: an orthogonal analysis, factor loading bivariate correlation between the observed variables and components are equal.

Step 4: Rotate the interpretation of solution Upheld the interpretation of the rotated solution is measured by each of the components, determine what is just. In short, it is to show that a given component variables to identify high loading, and to determine what is included in these variables have in common. Typically, a short name that describes its content is assigned to each retained component.

The first decision to be made at this stage loading is a factor that must be considered to decide how large is "large." Stevens discusses guidelines for testing the loading factor. The absolute value is greater than 0.40 then the principal component analysis is given that an introductory treatment, however, just a loading "big" idea.

Step 5: Create Factor scores Once the analysis gets complete, it stands on the subject retained components indicate where to assign scores for each subject is often desirable. For example, in the present study maintained two components give the financial help of a familiar component and were interpreted as a component. With this done, the component scores as predictor variables or the subsequent analysis can also be used as a criterion variable.

Step 6: A table summarizing the results Some articles publish the result of our analysis and it presents a table to prepare rotated factor pattern is generally desirable. Items include variables being analyzed responses to questionnaires, so it really questionnaire items in the table itself to reproduce may be helpful.

3.3.2 Linear Discriminant Analysis

Feature dimension reduction is a great approach for improving the performance of an activity recognition system that go for using visual features. In a typical classification problem the system designer chooses a number of features. The system designer believes that each of these features help in some discrimination But it is difficult to ensure that the information contained in each feature is extra to what is already available through the remaining feature At most if the new features do not contain any new information they will b e ignored. In practice unfortunately investigators of pattern classification problems have observed that beyond a certain point performance. The basic source of the and an increase in feature dimension parameters to be estimated with that limited amount of data Hughes et al. Inaccuracies

in parameter estimation process lead to degradation in recognition performance. Similar problem arises when very complex models are used to model the data.

Feature dimension reduction not only improves the recognizer performance but can also speed up the pattern recognition process Fisher introduced a technique of dimension reduction to a one dimensional linear subspace for the problem of two class classification Fisher et al. Other researchers later extended this technique to handle multiple classes Rao et al. [50]. I will call this technique by the name linear discriminant analysis LDA. LDA also called Fisher discriminant analysis or multiple discriminant analysis in literature is a widely used technique for reducing the feature dimension Duda and Hart et al. [51] This chapter gives a brief introduction to the technique of LDA and its application to activity recognition.

Two populations:

1. Separation

Suppose we have two populations. Let $\boldsymbol{X_1, X_2 \ldots, X_{n_1}}$ be the n_1 observations from population 1 and let $\boldsymbol{X_{n_1+1}, X_{n_2+1}, \ldots, X_{n_1+n}}$ be n_2 observations from population2. Note that, $\boldsymbol{X_1, X_2 \ldots X_{n_1}}$, $\boldsymbol{X_{n_1+1}, X_{n_2+1}, \ldots, X_{n_1+n}}$ are $p \times 1$ vectors. The Fisher's discriminant method is to project these $p \times 1$ vectors to the real values via a linear function $l(X) = a^t X$ and *try to separate the two populations as much as possible, where a is some* $p \times 1$ vector.

Fisher's discriminant method is as follows:

Find the vector $\hat{a}$ maximizing the separation function $|S(a)|$,

$$S(a) = \frac{\overline{Y_1} - \overline{Y_2}}{S_Y} \qquad (3.11)$$

Where $\quad \overline{Y}_1 = \frac{\sum_{i=1}^{n_1} Y_i}{n_1}, \overline{Y}_2 = \frac{\sum_{i=1}^{n_1} Y_i}{n_2}, S_Y^2 = \frac{\frac{\sum_{i=1}^{n_1} {Y_i}^2}{n_1} + \frac{\sum_{i=1}^{n_1+n_2} {Y_i}^2}{n_1}}{n_1+n_2-2}$

And $Y_i = a^t X_i \; for \; i = 1,2 \ldots n_1 + n_2$

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant features separate object or event which is the best linear combination of two or more sections to find ways to use statistics and machine learning. The resulting combination to reduce the dimensionality prior to classification, and more commonly used as a linear classifier, or can be. Here we amplitude decrease in activity for recognition only after applying PCA, LDA is used for classification.

3.4 Classification

3.4.1 Support Vector Machine

Support Vector Machine (SVM) in the first Colt-92 Boser, Guyon, and introduced by Vapnik in 1992 was heard. Support vector machines (SVMs) classification and regression [52] used to have a set of related supervised learning methods. They are a family of generalized linear classifiers. In other words, the support vector machine (SVM) automatically while avoiding over-fit the data to maximize predictive accuracy using machine learning theory predicts that a classification and regression tool. Support vector machine nips initially popular with the community and now is an active part of worldwide research in machine learning. When using pixel maps as input, SVM becomes famous; It is a handwriting recognition task [53] for detailed features sophisticated neural network gives comparable accuracy. It also specifically for pattern classification and regression-based applications, such as hand writing analysis much further, such as facial analysis and has been used for many applications. Support Vector Machine (SVM) foundation Vapnik [54] is developed by empirical and thus better performance has been achieved popularity for many promising features. Building used by conventional neural networks, the traditional empirical risk minimization (ERM) principle, is shown to be better, the structural risk minimization (SRM) principle uses. SRM ERM minimizes the error on the training data where the expected risk, but the overhead is low. It is the goal in statistical learning more and more with the ability to generalize this difference which equips SVM. SVMs have been developed to solve the classification problem, but recently they regression problems has been extended to solve.

Introduction to SVM:

Used for applications such as learning, supervised and unsupervised learning, for the first while working with neural networks showed the best results. MLP uses feed forward and recurrent networks. Multilayer perceptron (MLP) universal approximation properties include continuous nonlinear functions and input-output pattern learning and too much input and output [59] is involved with the advanced network architecture.

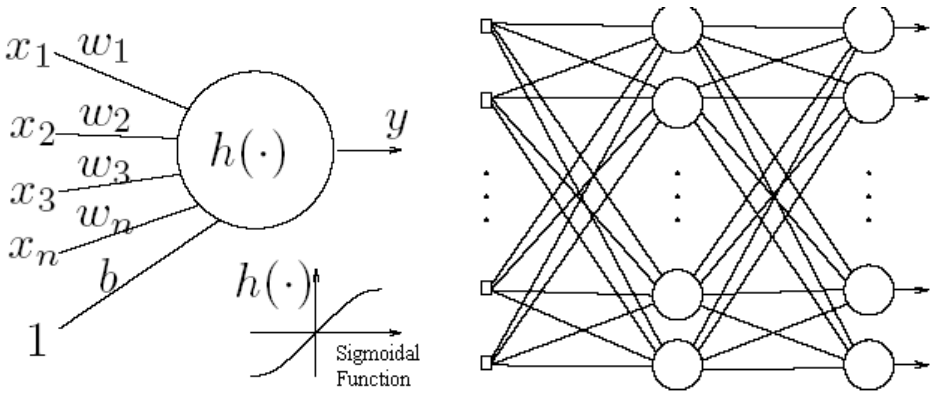

Fig 3.9 showing selection function of SVM

Problem still persists here. There are more than one local minima's are found. How many neurons are required his is also a big task and since this only gives the result of optimality. Another thing to consider is used neural network to converge to a solution, even if it is a unique solution may not be in the result. Now we plot the data we try to classify it and see another example where we can categorize see that many hyper plane. Question is which one gives better result?

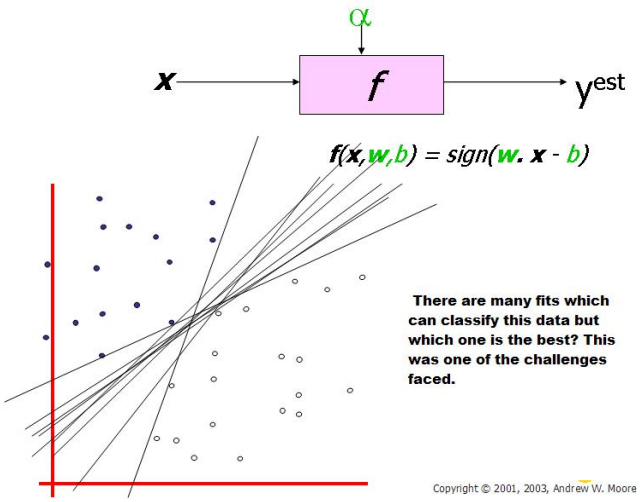

Fig 3.10: Many hyper planes showing classification

From the above illustration, the data is different that many linear classifiers (hyper planes) are. Although only one is able to provide maximum separation. only one is able to provide maximum separation and classification.. Since the hyper plane whichwhic we are going to use here may touch some class..and we don't want that as this may result into bad classification so concept of max margin emerges. The next illustration problem described above provides a solution that gives the maximum margin classifier instance.

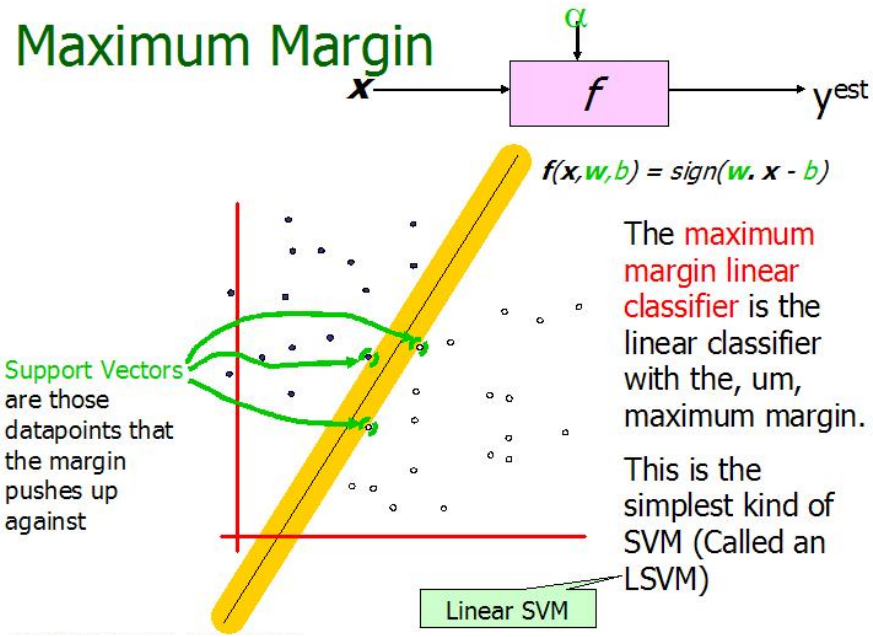

Fig 3.11: Showing maximum margin concept

Expression for Maximum margin is as follows.

$$margin = \arg\min d(x) = \arg\min \frac{|x.w+b|}{\sqrt{\sum_{i=1}^{d} {w_i}^2}} \quad (3.12)$$

Maximum linear classifier with the illustration above is the maximum limit. Here we are only showing a simple illustration of a simple example. Another intriguing question is why the maximum margin is required? Which include better empirical performance are some good explanations. Another advantage can be that we here can shun the local minima and better classification can be achieved. Let us go for representation of mathematical SVM and for this tutorial we try to present a linear SVM. With the goals of SVM hyper plane separating the data and use the kernel trick to extend the limits are non-linear [60]. We aim to compute the SVM to correctly classify all the data that is to see. We have to mathematical calculations,

[a] If Y_i= +1; $wx_i + b \geq 1$

[b] If Y_i= -1; $wx_i + b \leq 1$

[c] For all i, $y_i(w_i + b) \geq 1$

Here equation X is a feature vector point. W is the weight and it is also in form of a vector. Then [a] must always be greater than zero to separate the data. Among all the possible hyper planes, SVM hyper-plane distance is as large as possible, where one selects. A good training is required obviously and every test feature vector data from training vector is located at some distance. Now select the hyper-plane data [55] is far from possible. The maximum margin hyper plane of the journey to the nearest points on the convex hull lines between two datasets bisects.

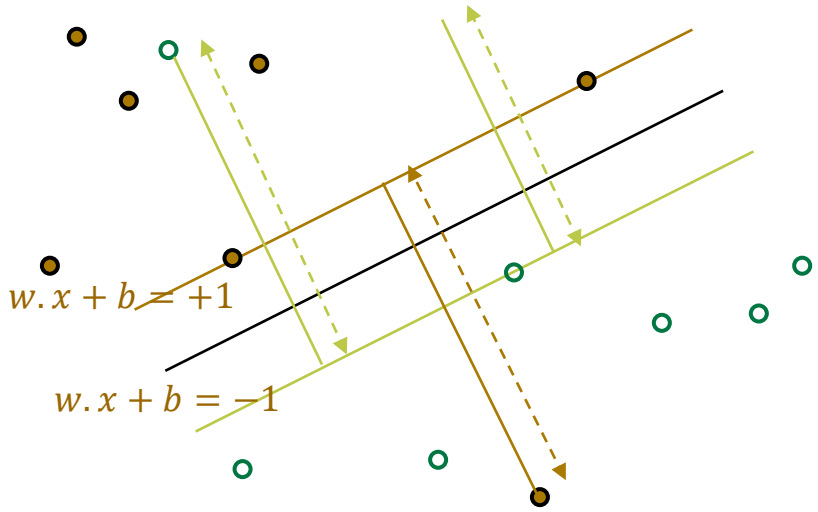

Fig 3.12: Representation of Hyper planes.

Original distance of the nearest point on the hyperplane x as hyper plane can be found by X-$_{max}$. Similar to the other side of the issue we have a similar situation. And we will resolve this issue by using two distance separating hyperplane obtained by subtracting the summed distance.

$$\text{Maximum Margin} = M = 2 / \|w\| \qquad (3.13)$$

Now maximization of the margin is same as finding minimum of the same equation [56]. For this we have to have a optimization problem and at the same time we have quadratic optimization problem and we need to solve for w and b.. The solution lies in constructing a dual problem and where a Langlier's multiplier α_i is related to it. We need to find linear constants which are important part of SVM (w and b) such that Φ (w) =½ |w'||w| is minimized;

$$\{(x_i, y_i)\}: yi\ (w.x_i\ +\ b)\ \geq\ 1 \qquad (3.14)$$

Now solving: we get that $w = \sum \alpha_i\ .x_i\ ; b =\ y_k - w.x_k$ For any $x_k\ such$ that $\alpha k \neq 0$

Now the classifying function will have the following form: $f(x) = \sum \alpha_i y_i x_i\ .x + b$

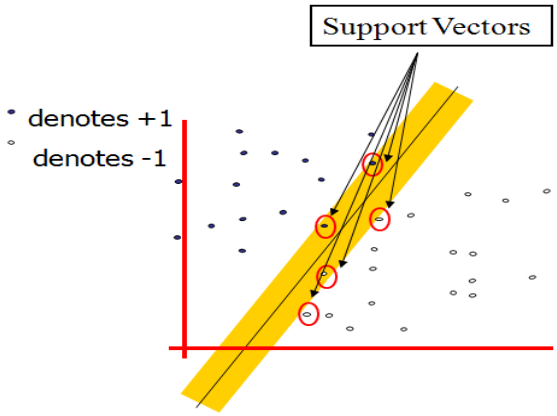

Fig 3.13: showing of Support Vectors

SVM Representation

In this we present the QP formulation for SVM classification [55]. This is a simple representation only.

SV classification:

$$\min \| f \|_K^2 + C \sum_{i=1}^{l} \xi_i \qquad y_i f(x_i) \geq 1 - \xi_i \; for \; all \; i \;\; \xi_i \geq 0 \qquad (3.15)$$

SVM classification, Dual formulation:

$$min \sum_{i=1}^{l} \xi_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \xi_i \xi_j y_i \, y_j K(x_i x_j) \;\; 0 \leq \alpha_i \leq C \; for \; all \; i; \qquad \sum_{i=1}^{l} \alpha_i y_i = 0$$

(3.16)

Variables ξ_i are called slack variables and they measure the error made at point $(\mathbf{x}_i, y_i)$. Training of SVM becomes a task in itself when number of features are more then either we uses a dimension reduction techniques like PCA or LDA or the fast SVM. Here we are showing methods for the fast SVM.

SVM for Classification

SVM is a most popular method for data classification. Neural networks, however, are sometimes considered unsatisfactory results than the ones that are easy to use. A classification task which usually consist of training and test data, some data

instances [65], is involved with. Each example in the training set contains one target prices and features. Only [66] trial has been set a target of SVM characteristics, a model which predicts target value of data instances is producing.

Supervised learning is an example of SVM classification. Famous labels help system or signal is not correctly performed. Validate the accuracy of the information system, or system to help you learn to correctly function to be used to point to a desired response. SVM classification includes a step known as the sections are connected to identity. This method is called feature extraction and some times selection. More of the times Feature extraction and SVM classification with unknown samples when an experiment is not necessary to predict. Some of the key squares procedures which they [65] can be used to identify differences that set.

3.3.2 NEAREST NEIGHBOR

Easiest known from the very beginning of the classification system emergence in the machine learning field is Nearest Neighbor classification technique – it works on to identify the nearest neighbors to a query based example and using these neighbors to find the class to which the belong. Poor run-time performance issues that computational power is available these days are not such a problem with this approach is of particular importance today for classification.

Intuition underlying nearest neighbor technique to classify the query base point and there class identification, is quite simple. Techniques more commonly used in determining class k nearest neighbors, where k nearest neighbor (k-NN) classification is known as it is often useful to take into account more than one neighbor. They should be in memory at run time, i.e. since training examples are required at run time, it is sometimes also called memory-based classification. Induction is delayed to run time. It is also called that no formal training is required . Training and testing is done at the same time because the classification or classification based case-based examples.

The basic idea of a two-dimensional feature space on a two-tier problem that Figure 3 depicts the nearest neighbor classifier is shown. Classification problem is

solved by simply majority voting and distance and weights used. This you can see below.

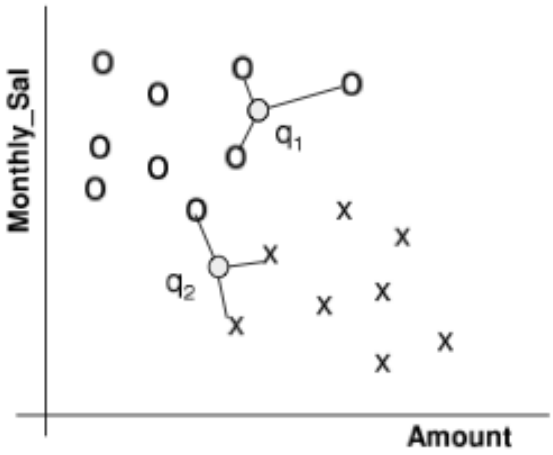

Fig 3.14: distribution and distance metric for KNN

The k-NN classification has two stages; (1) finding of best nearest neighbors and the (2) is the determination of the neighbors use the class.

KNN suffers severely from the nearest neighbor method called "dimensionality curse of." There are many aspects of this curse. Computationally, dimensionality M ≥ 20 data is very slow, as stated above. More easily, as the method increases the accuracy of the meter gets worse. K-nearest neighbor (KNN) classifier is a training sequence matrix between the current image sequence and calculates the distance. So why are not equal meaningful nearest neighbors, all points in a higher dimensional space are far away from each other. K-nearest neighbor (KNN) classifier is a training sequence matrix between the current image sequence and calculates the distance. Compare this to a large training set, computationally expensive to prove. All the action sequences in an alternative method can be calculated by taking the average of the mean and are used for all classes. Potential space and time performance, attitude and look of the image training data of features, image description and points the way depends on the distance between measured for that adapts.

In various domains of data in high dimensions can be used but for the KNN it is suggested that one should not use data which have larger dimension as the KNN is

not capable of doing the classification for a large data set. KNN classification can be used at different abstraction levels. Frame level one and level two whole scenes. Different lengths need to resolve the issue of the frame.

Curse of Dimensionality this used as the effectiveness of of the distance distribution plot inter-pair in the training set. / 2 pairs - N is large, all N (1 N) is enough to take a random sample. Delivery to the large spread is desirable to compare their means. If so, then we can say that dimension of the training set is small. Using a p-norm distance with p vectors that are real value for example <2 may be beneficial rather than Euclidean distance Agrawal et al., [22]. The distance function

$$d_p(x,y) = \| x - y \|_p = (\sum \| x - y \|^p)^{1/p} \quad (3.17)$$

The case p = 1 results in a distance known as Manhattan distance:

$$d_i(x,y) = \sum |\text{xi} - \text{yi}| \quad (3.18)$$

The case p = 0 is called Hamming distance:

$$d_0(\text{x},\text{y}) = \sum_{i=1}^{m} I(x \neq y) \quad (3.19)$$

Where I(α) is an indicator function that is either true or 0.

I (α) = 1 if α holds true, I(α) = 0 otherwise.

To give place to the variance of each dimension Mahalanobis distance is used. Rodriguez et al. [35] gives a method that generates spatio-temporal templates which effectively incorporates the intra-class variance at a place.

Bayes error rate for each instance on minimum average error probability is all examples. For example, the optimal prediction highest probability of any given X that is labeled x. For example, if the error probability is one minus the probability of the label.

This rate in actual can be written as

$$E^* = \int p(x)[1 - \max p(i - x)] \quad (3.20)$$

Here maximum can be find over any of the class $i = 1$ to $i = c$.

3.3.3 SVM in activity recognition

Since in practical applications situations are rare where only two classes are present so we have to have a method for multi classes, for it evolution of multiclass SVM took place by two popular and widely used approaches given by Hsu et al.[]. (1) Combines a number of two-class classification SVM in such a manner to make a multi-class classifier, while (2) Directly solves a multi-class classification problem with the training samples is given by Weston et al.[18]. Latter have decision-making function that is difficult to work and the training, testing and detection processes are too time consuming. So, the first method is in true sense more applicable for pragmatic situations. Several algorithms are invented including the one-against-all method, the one-against-each method and SVM-BTA.

One-against-others

It is one of the earliest methods of SVM multi-class classifier Wang et al. [34]. It requires N SVM classifiers for a problem of N-class classification. Here all of categories are considered while training each SVM structure, and the samples of a certain category have the true while the remaining have the false. Therefore, its complexity is high and speed is slow. Beside this, the discrepancy between the false and the true samples in the distinguishing with many classes there is a chance of degradation of recognition rate. Another problem with this method is that unclassified samples also remain there in it.

One-against-one

One-against-one method described by Krebel [68] needs classifiers, each of them is trained on various samples from any two classes. When all the classifiers got trained, a voting technique is used for testing. The false samples are assigned to the class that has largest number of votes. Hence the classification accuracy is improved in comparison to one-against-rest method,, however, there are still remains unclassified samples and without results.

3.3.4 KNN in activity recognition

KNN is a basic and from the beginning is used for the pattern recognition. It has been developed from the neural network and is a distance metric classification where different distance are used some of distance are as Euclidean and Mahalanobis distances. In activity recognition since there are many classes and we need a classifier, which gives result for a query, and tell us which class this belongs. Since KNN is based on query and we need only this type of classifier. In other classifiers like SVM or others we have only two class division for its basic form. We have to mold these classifiers but in KNN this is not required. It in its basic form is a multi class classifier. But the problem is of dimension KNN can not do justice with large dimension of data. So we can use KNN if we have used some dimension reduction technique before apply data to the KNN classifier.

CHAPTER 4

Experimental Results:

This Research is based on the activity videos of KTH. Since this is a method based on basic actions it can also be used for similar kind of dataset like Weizmann. KTH contains six basic activities e.g. ‘walking’, ‘running’, ‘hand-waving’, ‘handclapping’, ‘jogging’, ‘boxing’. Each action has 100 videos for four different scenarios in different light condition, indoor and outdoor conditions. Silhouette of human image abstraction from frames generated by the video is in itself a big task, since KTH data set is having different scenarios. Silhouette is abstracted and preprocessing is done to make it in binary form and too make all silhouettes of uniform dimension that is 25×25. All these images of 25×25 make a feature vector and all feature vector of different action videos make feature set. Experiments are done over these feature set to classify and analyze action classification. These experiments have reduced dimension feature set. This reduction is also matter of choice but here we have only used PCA. SVM, KNN and LDA classifications are what we have used here results and analyses are described further. The six fold cross validation process is used to evaluate accuracy of the system.

KTH dataset On the road with different clothes on the street S3 S1, S2 vary with scale road,: human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by four different scenarios Six of the 25 subjects thus with the existing video database and S4 indoors as illustrated below. The current database contains 2391 sequences. All sequences with a stationary camera with 25fps frame rate moved on homogeneous background. The spatial resolution of 160x120 pixels down-sampled sequences and the average length of four seconds.

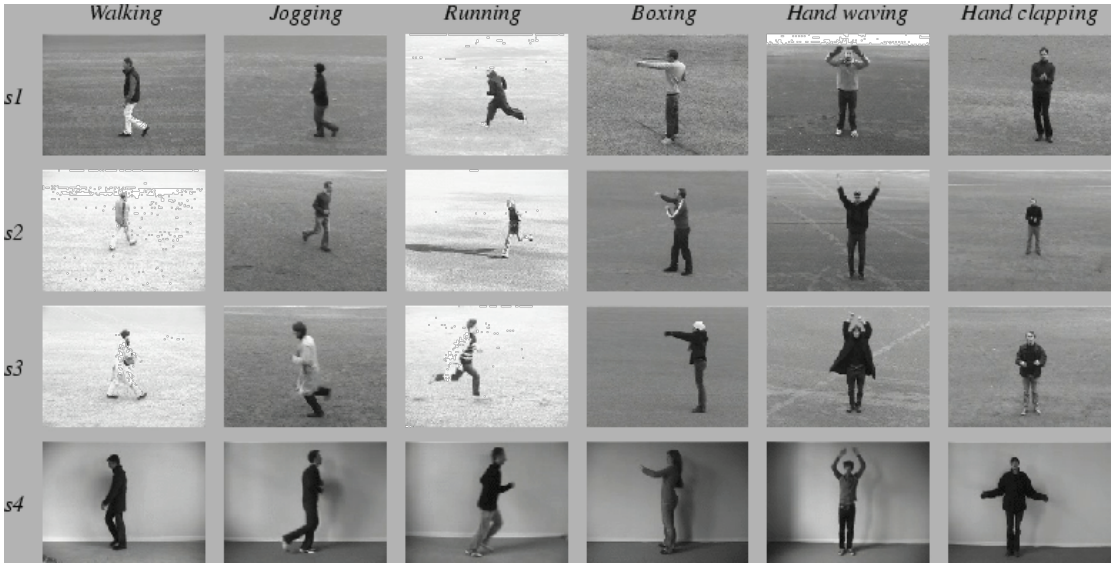

Fig 4.1 Samples of video set from KTH

4.1 Classification by LDA classifier after PCA

Here PCA is used for dimension reduction, since we are concerned about the speed of the process and this results into low speed of processing if we use kernel-based algorithms. PCA and LDA both are linear processors and this combination results into a faster processing and less complexity. We have described other linear discriminators and classifiers in this research but they are not as fast as this combination of PCA and LDA. But the speed comes at the cost of efficiency. This combination gives efficiency of 90 percent that is not enough to compete with the present available methods of classification. We have described other better methods in this research in further sections.

4.2 Classification by PCA and Multiclass SVM

PCA is used here for dimension reduction and Multiclass SVM is used for classification. SVM is of many to one classification. Training is done for one activity as one class and all others activities as second class. Such six structures are formed which are used for classification. Testing is also done one against other and the same structures are used for it. Speed of this process is also high since only linear processes are used but lags from PCA+LDA combination. Efficiency of process is also very high and it is 92 percent.

4.3 Classification by PCA and Nearest Neighbor

After reducing dimension of feature set by PCA nearest neighbor for classification is used. Efficiency of nearest neighbor is lower then that of multiclass SVM. But this is definitely better in terms of speed then SVM. This lies between LDA and SVM. Efficiency of this method is 92 percent. SVM and KNN are used in this research to reach a efficiency level of nowadays.

4.4 Classification by Nearest Neighbor and multiclass SVM on PCA

Both the multiclass SVM and nearest neighbor does not result into enough efficient method. Both this results in efficiency smaller than or equal to 92 percent. But if we use combination of both these methods this results in efficiency greater then 94 percent.

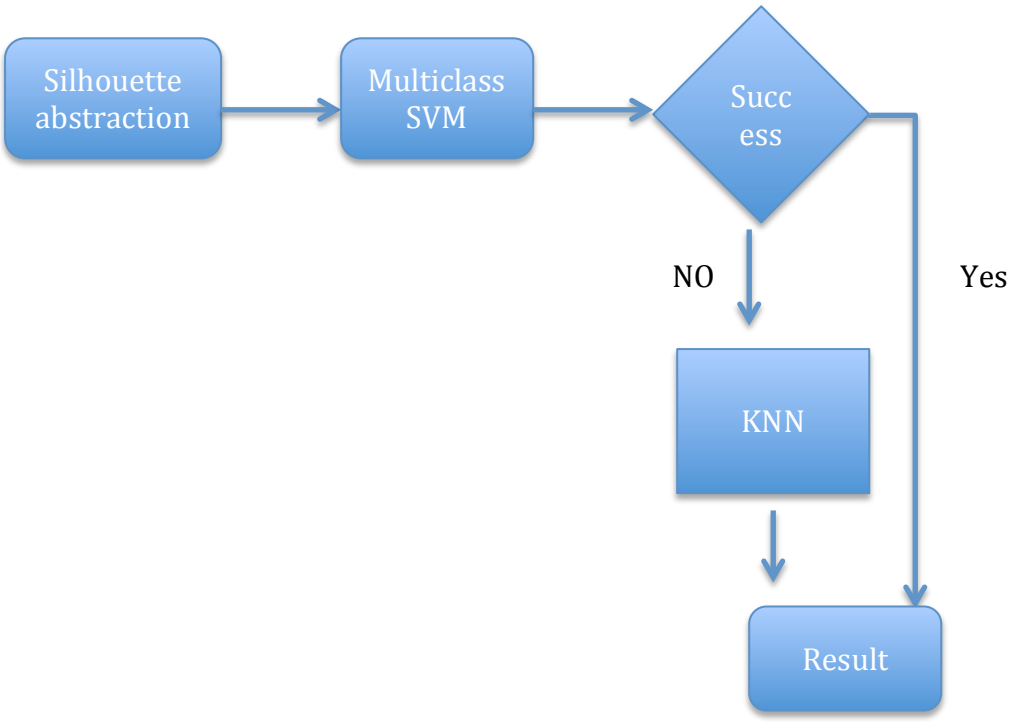

Fig 4.2 Showing algorithm used for NN+SVM

Table1: Average Recognition rate of all six activities for different methods in %

Activity	PCA+LDA	PCA+NN	PCA+SVM	PCA+NN+SVM
'Running'	80	91	92	95
'Walking'	84	90	91	94
'Boxing'	81	91	92.5	95
'Jumping'	85	89	90	93
'Hand-waving'	83	88	90	93
'Handclapping'	88	88	91	95

Since SVM used here is a multiclass and testing results depends on iteration. But iteration cannot be infinite and have to be a fixed number. In this case where one to many SVM is used SVM not always is successful in giving a result as described in section 3 that give details about multiclass SVM. When multiclass SVM is not able to work out KNN comes in and gives a result. Combining both of them results in a efficiency better then both of them. This efficiency is shown in table1 and fig4.1.

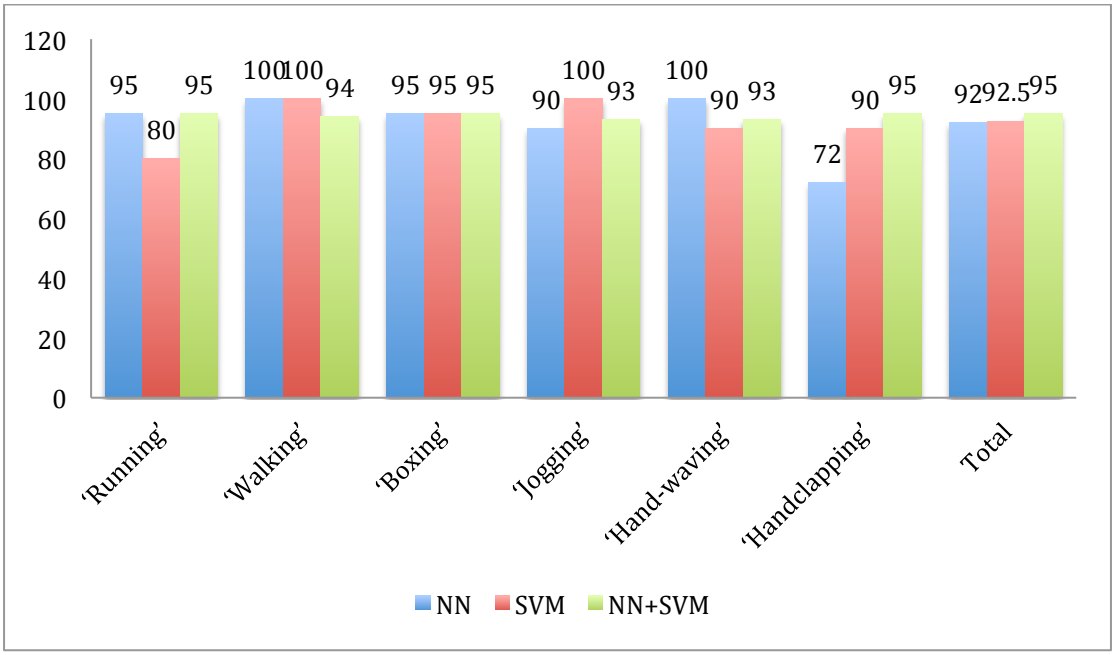

Fig 4.3 Bar Chart Showing Recognition Rate of Different activities in % on KTH data set

Fig 4.3 shows applications of different classification techniques that we have used here like SVM, KNN and LDA for classification. Since it may have cluttered the picture LDA is not shown here. The combined result is also shown here that is of KNN and SVM. Since SVM used here is a multiclass and testing results depends on iteration. But iteration cannot be infinite and have to be a fixed number. In this case where one to many SVM is used SVM not always is successful in giving a result as described in section 3.5.1 that give details about multiclass SVM.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This research is applied on the KTH dataset as various other authors have done activity recognition work on the dataset, which is easily available. Some authors have also prepared their own data set but in our case we have used a popular an accredited dataset so there remains no hidden thing in this research and validity of the research can be accepted widely.

Here a method is developed in this research, which is based on the three different classification techniques in basic part of it and using two classification techniques develops one technique of classification.

5.1 Proposed Method

Here research is done over the video dataset. It is used for action recognition by various authors. Linear discrimination analysis is done over the generated feature vectors and this gives a result that is better than the authors who has given the basic element silhouette in their research. But this at the same time do not give a result which is up to mark for the new computational technology and is not regarded as the modern classification technique. Silhouette information that is a shape feature of the action and activity is used.

Silhouettes are extracted by foreground segmentation. Silhouette is further divided into fixed number of cells from which features are extracted easily and with less computation. These features are arranged in a manner to preserve the time feature of an action. Since large number of frames and hence large number of features therefore a dimension reduction technique is applied. These features then used for training the Support Vector Machine and identification is done with SVM and K-Nearest Neighbor. Analysis of these combination e.g. PCA, SVM and KNN is done in this paper in the further parts of it. Unlike others here two Classifiers in series, which results into better efficiency have been used. The goal of this research was to find a

temporal segmentation of recordings obtained from human activities. This was performed in the wider context of activity classification. Currently, many algorithms obtain an implicit segmentation as a side effect of direct activity classification. In this research the primary goal was finding a temporal segmentation, which is assumed to be able to aid the classification step. This problem is reduced to finding change points in temporal data sets originating from inertial sensors.

To test our proposed method based on SVM and KNN. Applying individually these methods does not result in good efficient result. But when we are applying these methods in complimentary way it results in an efficient method as described in section 4. In comparison to other methods as shown in table 2 our method proves to be better than many methods presented by various authors. It have compared results with various authors that are applicable over KTH data set. KTH dataset has been chosen because of in different scenarios and requirement of a robust method for such complex situations. And it also contains basic activities that can be elementary unit for complex activities.

5.2 Contributions

In this research texture based segmentation of the human blob is used that is totally a new concept for the activity recognition. Texture based segmentation is rarely used by other authors and they use some complex methods or the very basic background subtraction techniques but here it has been shown that the it works perfectly if we can used texture segmentation for human blob detection and its recognition. This research is for the very basic part of silhouette recognition and based on that activity recognition. Without these shape descriptors activity recognition can not be done and we have to have a shape or temporal descriptor for activity recognition.

The second new thing done in this paper is using a classification technique that is based on two basic classification techniques. Here KNN and SVM is used simultaneously. It is proved that the two techniques are totally different and complement to each other. If it is possible to use best of these two techniques it results into a very efficient classifier.

Over described techniques are tested for the KTH data set and results are included in this literature.

5.3 Limitations

There is stillroom for improvement. This technique generates a lot of data that is redundant although, dimension reduction technique is used for it. For a better and more robust analysis of the algorithm it can be applied to other real-world data sets. It is not used for real time data. And we need to develop this technique for real data. Since real data requires more fast process and so the less dimensional features must be extracted from it. Linear discrimination analysis is done over the generated feature vectors and this gives a result that is better than the authors who has given the basic element silhouette in their research. But this at the same time does not give a result that is up to mark for the new computational technology and is not regarded as the modern classification technique. Silhouette information that is a shape feature of the action and activity is used.

5.4 Future research

This research have used two tot two popular techniques one is texture based segmentation for shape descriptor based recognition. Texture based technique is used to get out the silhouette from a frame of action recognition video. Texture based segmentation is working well here on the data set available at hand but on realistic situations they may fail as these situations have clutter and complex background. There is need of thorough research on segmentation by differentiating between more than two textures as complex situations have multiple textures and this may not work for the scenarios that have more than two textures.

Here we have described activity recognition by using two classifiers used at same time. Results show that this method is more robust than the techniques, which use only one classifier. Future research can be based upon more than two classifications. The classifiers should be selected in such a way that one compensate the errors done by the other classifier. More classifier helps in more complex conditions. As the research is applied to realistic situations complexity increases as to recognize these types of actions one do need more robust classifier. In some but not

all situations one classifier can work well. For more efficiency definitely one needs more robust classifier and this can be done by combination of more than one classifier.

Future demands research on combination of classifiers that can work well with the action recognition. Since any random combination cannot be used. Two classifiers, which have nearly similar techniques, must not be combined together as they will not be useful as they both are doing classification on same basic concept.

References

[1] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, Comput. Vis.Pattern Recogn. 981 (2005) 984–989.

[2] Cen Rao, Alper Yilmaz, Mubarak Shah, View-invariant representation and recognition of actions, International Journal of Computer Vision (IJCV) 50 (2) (2002) 203–226.

[3] S. Ben-ari, D. Xu, B. Zhang, H.J. Zhang, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.

[4] X. Wu, Y. Jia, for Pfinder W. Liang, Incremental discriminant-analysis of canonical correlations for action recognition, Pattern Recognition 43 (12) (2010) 4190–4197

[5] T. Olson, and F. Brill, "Moving object detection and event recognition algorithms for smart cameras," DARPA Image Understanding Workshop, pp. 159–175, 1997.

[6] Kumar, S., Raman, B., & Sukavanam, N. (2011). Human action recognition in a wide and complex environment. SPIE –The International Society for Optical Engineering, 7871.

[7] R. Xiao, W. Li, Y. Tian, X. Tang, Joint boosting feature selection for robust face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1415–1422.

[8] Y. Wang, G. Mori, Max-margin hidden conditional random fields for human action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 872–879. [10] M. Ahmad, S.-W. Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (2008) 2237–2252.

[9] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82– 98, Jan 1999.

[10] M. Brand and K. Huang. A unifying theorem for spectral embedding and clustering. In *Proc. of the Ninth International Workshop on AI and Statistics*, 2003

[11] Tseng X., 2013. Intelligent multi-camera video surveillance: a review. Pattern Recognit. Lett. 34, 3–19, Extracting Semantics from Multi-Spectrum Video.

[12] T.H. Thi, J. Zhang, L. Cheng, L. Wang, S. Satoh, for NCDAV, Human action recognition and localization in video using structured learning of local space-time features, in: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010.

[13] Uddin, M.Z.; Kim, D.H.; Kim, J.T.; Kim, T.S. An indoor human activity recognition system for smart home using local binary pattern features with hidden markov models. *Indoor Built Environ.* 2013, *22*, 289–298.

[14] M. Abdelker, G. Shaogang, X. Tao, Recognizing action as clouds of space-time interest points, Comput. Vis. Pattern Recogn. (2009) 1948–1955.

[15] S. Red, O. Oreifej, M. Shah, Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories, in ICC11, 2011.

[16] D. Weinland, Edmond Boyer, Action recognition using exemplar-based embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008

[17] R. Poppe, M. Poel, Discriminative human action recognition using pairwise CSP classifiers, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2008, pp. 1–6.

[18] T.B. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, Comput. Vis. Image Underst. 104 (2006) 90–126.

[19] Hu, M.; Wang, Y.; Zhang, Z.; Zhang, D.; Little, J.J. Incremental learning for video-based gait recognition with LBP flow. *IEEE Trans. Cybern.* 2013, *43*, 77–89.

[20] A. Sanin, C. Harandi & B.C. Lovell. Spatio-temporal covariance descriptors for

action and gesture recognition. In Proceedings of the IEEE International Workshop on Applications of Computer Vision (WACV), Tampa, FL, USA, 17–18 January 2013; pp. 103–110.

[21] E. Gavrilla, , Y. Zhao & W.Xiong, (2010). Active energy image plus 2DLPP for gait recognition. Signal Processing, 90, 2295–2302.

[22] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (1999) 428–440.

[24] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transaction on Circuits and Systems for Video Technology 18 (11) (2008) 1473–1488.

[25] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[26] K. Toyama and A. Blake. Probabilistic tracking in a metricspace. In *ICCV*, pages 50–59, 2001.

[27] R. Polana, & R. Nelson (1992). Recognition of motion from temporal texture. In Proc. computer vision and pattern recognition 1992 (pp. 129–134).

[28] J.W. Davis, A. Tyagi, Minimal-latency human action recognition using reliable-inference, Image Vis. Comput. 24 (2006) 455–472.

[29] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Transaction on PAMI 23 (3) (2001) 257–267.

[30] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-timeshapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 2247–2253.

[31] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, Comput. Vis. Pattern Recogn. (2008) 1–8.

[32] Polana, & Nelson (1992). Recognition of motion from temporal texture. In Proc. computer vision and pattern recognition 1992 (pp. 129–134).

[33] A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: IEEE International Conference on Computer Vision, vol. 2, 2003, pp. 726–733.

[34] A. Fathi, G. Mori, 2008. Action recognition by learning mid-level motion features. In: IEEE Conf. Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8.

[35] J. Acevedo-Rodr´ıguez, S. Maldonado-Bascón, R. López-Sastre, P. Gil-Jiménez, A. Fernández-Caballero, Clustering of trajectories in video surveillance using growing neural gas, Lecture Notes in Computer Science 6686 (2011) 461–470.

[36] Lu, J., Montemayor, A., Pantrigo, J., Sánchez, A., 2011. Human action recognition based on tracking features. In: Ferrández, J., Álvarez Sánchez, J., de la Paz, F., Toledo, F. (Eds.), Foundations on Natural and Artificial Computation,Lecture Notes in Computer Science, 6686, pp. 471–480.

[37] Ke, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B., 20011. Human action recognitionusing multiple views: a comparative perspective on recent developments. In: Proc. of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding. ACM, New York, NY, USA, pp. 47–52.

[38] Y. Gool, G. Xu, S. Tsuji, Understanding human motion patterns, in: International Conference on Image Analysis and Pattern Recognition, 1994, pp. 325–329.

[39] Los, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Systems Man Cybernet. 34, 334–352.

[40] L. Li, W. Wong, I.Y.-H. Gu, T. Qi, Statistical modeling of complex backgrounds for foreground object detection, Image Process. 13 (2004) 1459–1472.

[41] I. Laptev, M. Lindenberg, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8

[42] Wren, C. R., Azarbayejani, Dollar A., & Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and

Machine Intelligence, 19, 780–785.

[43] Y. Sivic, H. Foroosh, View-invariant action recognition using fundamental ratios, Comput. Vis. Pattern Recogn. (2008) 1–6.

[44] X. Hen, P. Niyogi, Locality preserving projections, Advances in Neural Information Processing Systems 16 (2003) 152–160.

[45] Niebels, J., Ohya, J., & Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. Computer Vision and Pattern Recognition, 379–385.

[46]J. Haralick, A. Yao, N. Razavi, L. Van Gool, V. Lempitsky, Hough forests for object detection, tracking, and action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 2188–2202. 8.

[47] R. Turk, X. Sun, H. Yao, P. Xu, T. Liu, Attention-driven action retrieval with DTW- based 3d descriptor matching, in: ACM International Conference on Multimedia, ACM, Vancouver, British Columbia, Canada, 2008, pp. 619–622.

[48] R. Kaiser, H. Yao, X. Sun, Actor-independent action search using spatiotemporal vocabulary with appearance hashing, Pattern Recogn. 44 (2011) 624–638.

[49] H. Scree, P. Milanfar, Action recognition from one example, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 867–882.

[50] M. Rao, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, S.A. Velastin, Recognizing human actions using silhouette-based hmm, in: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS ’09), 2009, pp. 43–48.

[51] R. O. Duda, P. E. Hart and D. G. Stork, Pattern classification (2nd Edition), Wiley-Interscience, 2000

[52] Vladimir Vapnik. Pattern recognition using generalized portrait method. Automation and remote control, 24:774–780, 1963.

[53] H. Zhou, P. Miller, J. Zhang, Age classification using radon transform and entropy based scaling SVM, in: Proceeding of the British Machine Vision Conference, 2011, pp. 28.1–28.12.

[54] V. Vapnik, R. Ronfard & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding, 115, 224–241.

[55] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, IEEE International Conference on Pattern Recognition, vol. 3, 2004, pp. 32–36.

[56] A. Elgammal, C.S. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 681–688.

[57] X. He, P. Niyogi, Locality preserving projections, Advances in Neural Information Processing Systems 16 (2003) 152–160.

[58] D. Cai, X. He, K. Zhou, J. Han, H. Bao, Locality sensitive discriminant analysis, in: International Joint Conferences on Artificail Intelligence, 2007, pp. 708–713. [35] R. Wang, X. Chen, Manifold discriminant analysis, in: IEEE Conference onComputer Vision and Pattern Recognition, 2009, pp. 429–436.

[59] L.K. Jia, D.Y. Yeung, Human action recognition using local spatio-temporal discriminant embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[60] L. Wang, D. Suter, Visual learning and recognition of sequential data manifolds with applications to human movement analysis, Computer Vision and Image Understanding 110 (2) (2008) 152–172.

[61] Y. Sheikh, M. Shah, M. Shah, Exploring the space of a human action, ICCV (2005) 144–149.

[62] Chaudhry, Avinash Ravichandran G. Hager, R. Vidal, Histograms of oriented

optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1932–1939.

[63] X. Wu, Y. Jia, W. Liang, Incremental discriminant-analysis of canonical correlations for action recognition, Pattern Recognition 43 (12) (2010) 4190–4197.

[64] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transactions on Circuits and Systems for Video Technology, 1051-8215 18 (11) (2008) 1473–1488.

[65] M.S. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1036–1043. <www.scopus.com>.

[66] M.K. Hu, Visual pattern recognition by moment invariants, IRE Transaction on Information Theory 8 (1962) 179–187.