ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Silhouette based human action recognition using PCA and ISOMAP

Jyotsna E[1], Akhil P V[2], Arun Kumar[3]

PG Student, Computer Science and Information System, FISAT, Angamaly, India[1]

PG Student, Computer Science and Information System, FISAT, Angamaly, India[2]

Assistant Professor, Computer Science, FISAT, Angamaly, India[3]

Abstract: Human action recognition in computer vision has attracted strong research interest in recent years because of its promising applications. Automated detection of a person's atomic movements is called human action recognition. Diverse methods have been proposed for detecting the human action due to the dynamics in the underlying representation of human body. This paper provides a glance over the significant researches conducted in this area and a comprehensive survey of the different taxonomies suggested by different researchers. A system is proposed for efficient action recognition based on the concept 'bag of correlated poses' by taking into account of the correlation between sequential poses in an action on both linear and non linear feature vectors.

Keywords: Action recognition, Correlogram of body poses, PCA, ICA, ISOMAP.

I. INTRODUCTION

Visual analysis of human behaviour has attracted a great deal of attention in the field of computer vision because of the wide variety of potential applications. Human behaviour can be segmented into atomic actions, each one indicates a single, basic movement. Recognizing human action [1, 2] is a key component in many computer vision applications, such as in video surveillance, human-computer interaction HCI, proactive computing, mixed reality collaboration and robot learning. To reduce human intervention in the analysis of human behaviour, unsupervised learning may be more suitable than supervised learning. However, the complex nature of human behaviour analysis makes unsupervised learning a challenging task.

The action representation and recognition is relatively old concept yet still immature. Although the researchers are investing enormous efforts to propose their approaches for action or activity recognition—the reality is that this field is still not applicable in many important areas. Therefore, this project concentrate on the major classifications on appearance based action representation. It is obvious that without having robust and smarter representations. So it will be difficult to recognize actions in a reasonable manner.

The most common methods applied for describing image cues are silhouettes and contours, motion optical flows, colour and texture, depth maps. Silhouettes, as well as edges and contours, are used to fit human body in images because the most of the body pose information remains in its silhouette. However, methods using edges rely on a background subtraction stage because of the difficulty of extracting human silhouettes in complex scenarios. Recently, depth cues have been included in several human pose recognition systems because of the depth maps provided by the multi-sensor KinectTM. This new depth representation offers near 3D information from a cheap sensor synchronized with RGB data.

Majority of the human action recognition works follow two paths. The first strategies involve initially locating the object of interest e.g. a human, tracking it so that a description of how the object changes over time can be formed, and then finally classifying the action. Many of the attempts using this approach are tested and evaluated on unchallenging datasets such as KTH which features a single person performing a single action in uncluttered setting. While tracking an object to differentiate one activity from another is not practical in cluttered scene. Tracking through motion or foreground segmentation can be sensitive and degrade ungracefully when errors occur in contiguous frames within a video sequence.

The other strategy avoids the drawbacks of tracking methods by directly analysing the motion patterns within the entire frame throughout the video sequence. Here methods consider such as video sequences, computing temporal templates constituting two components; a motion-energy image (MEI) and a motion-history image (MHI). The MEI is a binary representation of the motion occurring between sequences of frames while the MHI is a grey scale intensity representation where the most recent changes are lightest. It is considered to be a better idea for action recognition as it

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

captures the way in which the shape and motion evolve over time.

In the second path single layered approaches are quite common where an activity is considered as a particular class of image sequence and recognition is performed over an unknown input by categorizing it into its class. Human Activity Analysis [16] generalizes the methods for single layered approaches. One of the classifications is space – time approach and the other is sequential approach. Space-time approaches are those that recognize human activities by analysing the space time volumes of activity videos. And most of the activity recognition methods use spatial-temporal feature descriptors. Space-time approaches model a human activity as a particular 3-D volume in a space-time dimension or a set of features are extracted from the volume. Then volumes for video frames are constructed by concatenating image frames along a time axis. This is used for further similarity measures. On the other hand, sequential approaches treat human actions as a sequence of particular observations. More specifically, they represent a particular human action as a set of feature vectors or descriptors extracted from images and recognize activities by searching for such a sequence.

The spatio-temporal local feature-based approaches are getting an increasing amount of attention due to the reliability of the algorithms under noise and illumination changes. But these approaches, however, can only produce impressive results under conditions that the datasets are relatively less challenging in either limited viewpoint changes or uncluttered backgrounds. In addition to that, the performance of feature distance-based classifiers, such as KNN, would degrade when dealing with data that are composed of highly independent subsets, which can come from the multi-sensor data fusion phase.

For all these methods Dimensionality reduction technique is necessary. Data originating from the real world is often difficult to understand because of its high dimensionality. The resolution of the images may be 600x400 pixels, which means that the input data has 240000 dimensions. It stands to reason that most of the classification methods suffer and even fail in their goals when dealing with such kind of data due to their sensitivity to the dimensionality of the input data. Obviously, that data becomes intractable from the computational point of view when long image sequences are used and a dimensionality reduction technique is needed.

Once the human activities have been demonstrated and recorded, a dimensionality reduction technique may be needed in order to find a lower dimensional representation of the data. Both, linear (such as PCA and ICA) and non-linear (such as Isomap) dimensionality reduction techniques can be applied to reduce the dimensionality of the data. The choice of one or the other depends on the intrinsic nature of the data set. Reduce the dimensionality of the data can be viewed as a pre-processing step for classification purposes. This paper is a comparative study on the dimensionality reduction techniques implemented on the new Bag-of-correlated poses technique.

The rest of the paper is structured as follows. Section 2 summarizes important works related to action recognition. Section 3 includes the proposed method, and the details of implementation are described in Section 4. Simulation results and discussions of the proposed method are given in Section 5 and Section 6 concludes the paper.

II. RELATED WORKS

Various automated methods are developed for human action recognition. Bobick and Davis introduced [1] the new view-based approach to the representation and recognition of human movement. These approaches only stack the foreground regions of a person (i.e., silhouettes) to track shape changes explicitly. This method presents the representation of images using temporal templates – the first value is motion-energy image (MEI) which indicates the presence of motion and the second value is motion-history image (MHI) which is a function of the motion in a sequence. After computing the various scaled MHIs and MEIs, proposed method compute the Hu moments for each image and next check the Mahalanobis distance of the MEI parameters against the known view/movement pairs. Any movement found to be within a threshold distance is labelled with action. This is considered to be an effective method of representing and recognizing motion.

The paper [2] represent actions as three-dimensional shapes induced by the silhouettes in the space-time volume. This method chooses a Poisson-based descriptor because it reflects more global properties of the silhouette, and allows easy extraction of many useful shape properties. The Poisson descriptor tries to represent a shape by describing its silhouettes. The proposed method has a number of advantages like: - it is potential to cope with low quality video data, method does not require video alignment, linear in the number of space-time points in the shape, and the overall processing time of method takes less than 30 seconds. This approach can also be applied with very little change to general 3D shapes representation and matching. This method is fast and robust, because it contains both the spatial information about the human pose, and the dynamic information such as global body motion and motion of the limbs relative to the body.

The local descriptors and holistic features emphasize different aspects of actions and are suitable for the different types of action databases. In this paper [3] a unified action recognition framework fusing local descriptors and holistic features is proposed. Local descriptors like 2D and 3D SIFT feature descriptors based on 2D SIFT points are extracted and holistic features extracted with Zernike moments. The fusion approach adopted here gives a comparable result. In the previous work, most local descriptors approaches use the

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

spatiotemporal gradient information to extract interest points and most holistic features are based on the silhouettes or tracking. This paper introduces frame differencing to extract both local descriptors and holistic features and then use a bag-of-words approach to compute feature vectors for the feature fusion.

This paper [4] introduces a novel representation for human actions using Correlogram of Body Poses (CBP) which takes advantage of both the probabilistic distribution and the temporal relationship of human poses. Silhouettes are relatively easy to be obtained and it contains discriminative shape information, which is invariant to person's gender, body size, lighting condition, clothing, and appearance. Normalized silhouettes are used as input for system since body poses are encoded by silhouettes, which are robust to different clothing, appearance and illumination changes and also it saves the computation of feature description. This method is robust to unreliability of the segmentation, noisy labelling in training samples, and the speed of the action.

An extension of Correlogram based pose detection is done on [5]. Multi-view action recognition using local similarity random forests classifier and multi-sensor fusion is proposed in this paper. Here the idea of decision forests and Correlogram of body pose silhouettes as feature descriptors is fused with different camera views on the IXMAS silhouette dataset. Such a fusion method would benefit the recognition accuracy by describing the actions with more features. The problems like computational complexity for clustering and dimensionality reduction can be solved with the randomized forests classifier. Deterioration due to the unequal performance of each camera with respect to different actions is tackled by a new voting strategy. The random forests method has better efficiency and effectiveness than other learning algorithms like k-means especially when dealing with large scale data, and it can avoid the over-fitting problem by setting more decision trees.

In this paper [6] a shape-based feature descriptor Pyramid Correlogram of Oriented Gradients (PCOG) which is calculated from the Motion Energy Images (MEI) and Motion History Images (MHI) is used. By using the local and spatial layout properties, the PCOG descriptor captures the essential information of human actions and provides good discriminative for classification. Here a new Human action detection framework is proposed which target the human from the input by matching extracted HOG descriptors with the prototypical action primitives. The obtained vectors are used for periodical action partitioning. The output only contains the region of interest (ROI). Once a complete exercise cycle is extracted, two key frames and their corresponding MHI and MEI are selected to encode this movement. Then the action class is predicted by classifying the extracted feature descriptors (PCOG) using the trained classifier using the multi-class Support Vector Machine (SVM) with the RBF kernel. This approach can detect different action classes from a video Sequence. An indoor environment, like a gym is used for performance evaluation.

It is observed that sparse representations based on detected interest points, suffer from the loss of structure information. This paper [7] proposes a model which takes the motion and structure information simultaneously and integrates them in a unified framework and therefore provides an informative and compact representation of human actions. By applying the motion template to the volume with difference of frames (DoF), the motion information is encoded into the motion feature map (the motion history image), and structure feature maps are obtained from the structure planes extracted from the DoF volume. Two dimensional Gaussian pyramid and center-surround operation are performed on each feature map, in order to decompose feature maps into sub-band images localized in multiple center spatial frequencies. Then biologically inspired features are extracted using a two-stage feature extraction step ie; Gabor filtering and max pooling. The effectiveness of this method is evaluated on the KTH, the multi-view IXMAS, and the UCF sports datasets.

Paper [8] proposes an automatic video annotation algorithm by integrating semi-supervised learning and shared structure analysis into a joint framework for human action recognition. A new Semi-supervised Feature Correlation Mining (SFCM) method is introduced which leverages shared structural analysis for action recognition. Input Training Action Videos may contain both labeled videos and unlabeled videos. Features extraction by Harris3D and HoG/HoF BoW representation is performed for both training and testing videos to represent them. According to the distribution of the visual features, a graph model is constructed in training. Building upon the graph, virtual labels of the unlabeled data can be generated, during which shared structural analysis is applied to uncover the feature correlations to make the results more reliable. In this way, a classifier is trained for action recognition. To evaluate performances of proposed algorithm a brief comparison is made with SVM with 2 kernel, Bayes Optimal Kernel Discriminant Analysis (BKDA), Taylor-Boost (T-Boost) and Semi-supervised Discriminant Analysis (SDA).

Agarwal and Triggs describe [10] an approach for recovering 3D human body pose from single images and monocular image sequences. It detect human pose by direct nonlinear regression against shape descriptor vectors extracted from silhouettes where a silhouette shape is encoded by histogram of-shape-contexts descriptors. A sparse Bayesian approach is proposed based on nonlinear regression algorithm called Relevance Vector Machine (RVM). RVM's have been used to build kernel regressors for 2D displacement updates in correlation-based patch tracking. The main attraction of this method is it does not require a 3D body model or labelling of image positions.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

And the method is easily adaptable to different people or appearances. This method show promising results, being about three times more accurate than nearest neighbour methods.

The same approach has the different performance on the different database. This is due to the different characteristics of these databases. Some dataset has a larger data scale, different scenarios, changing backgrounds due to the camera zoom, more persons performing a particular action, and more intra-class dissimilarity in the shape of figures. Others may have a much lower data scale, only one scenario, static background, more action classes and more inter-class similarity in the local motion. Following section gives brief introduction to the datasets used in this for human action recognition which can be considered as benchmarks.

A. Weizmann Dataset

The Weizmann database contains classification dataset and robustness dataset. The Weizmann classification dataset is for action training and classification. The Weizmann robustness dataset is for testing the robustness of a human recognition method and contains two data subsets. It contains 90 low-resolution (180 × 144) video sequences from nine people, each performing ten natural actions: run, walk, skip, jumping jack, bend, jump in place on two legs, galloping sideways, wave one hand and wave two hands. All the videos are captured from a fixed viewpoint.

B. IXMAS Dataset

Above two sets were recorded in controlled and simplified settings. The first realistic-action dataset collected from movies and annotated from movie scripts is made in INRIA. INRIA Xmas Motion Acquisition Sequences (IXMAS). Multiview dataset used for view-invariant human action recognition. IXMAS aim to form a dataset comparable to the current state-of-the-art" in action recognition. It contains 11 actions, each performed 3 times by 10 actors of 5 males or 5 females. The actors freely change their orientation for each acquisition in order to demonstrate the view-invariance and no further indications on how to perform the actions beside the labels. Experiments were conducted using Leave-One-Actor-Out (LOAO) testing strategy.

III. PROPOSED METHODOLOGY

Figure 1 shows basic steps and approaches of human action recognition. Video of single activity is processed to generate the sequence of the images. Sequences of images are used for the HAR process. From sequence images silhouette image are generated. From every frame of data set, "Region of Interest" (ROI) containing the binary activity shape is extracted. Every video clip consists of single human activity. Silhouette images can be depth or binary.

Feature extraction is a special form of dimensionality reduction. The input data is transformed into the set of necessary feature vector is called feature extraction. From the extracted features, human activities are identified. Principal Component Analysis (PCA) is the popular method for the feature extraction. Another method is Independent Component Analysis (ICA). Linear discriminate analysis (LDA) is the classification tool that can be applied on PC and IC feature in order to get better result. This work also includes an unsupervised non-linear method called ISOMAP. Generally two vector quantization algorithms are used: namely ordinary K-means clustering and Linde, Buzo, and Gray (LBG)s clustering algorithm. Finally classification is done using the classifiers like SVM, multiSVM or HMM.

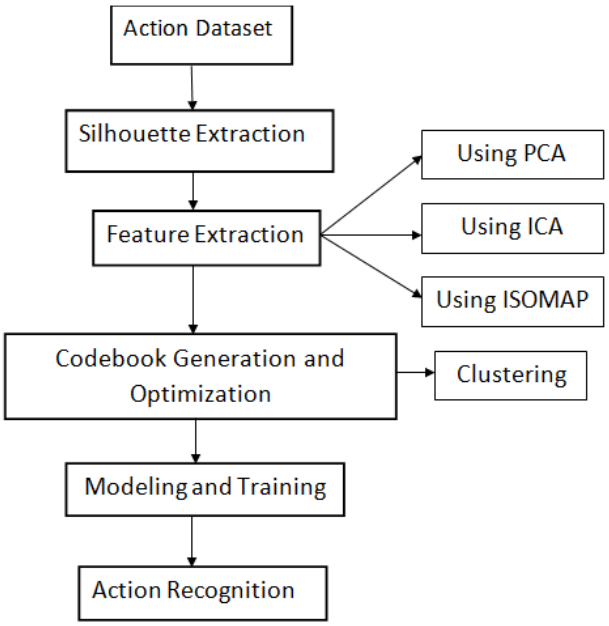

Fig. 1 Overview of the proposed system

A. Bag-of-correlated poses

The bag-of-features approach is a well-known method for action recognition. The bag-of-features-based approaches can be applied in classification by employing features as words. The Bag-of-Features representation is typically a normalized histogram, where each bin in the histogram is the number of features assigned to a particular code divided by the total number of features in the video clip. Due to its popularity, researchers are extensively considering this framework for their researches. It has similar versions as

- Bag-of-Feature
- Bag-of-Words
- Bag-of-Visual-Words
- Bag-of-Vocabularies
- Bag-of-Video-Words
- Bag-of-Points

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

The traditional bag-of-features representation disregards structural information among the visual words. If the codebook becomes very large, it may produce lower recognition. To encode the structural information Correlogram of human poses in an action sequence is introduced in this work[4].

Correlogram is the graphical representation of autocorrelation. Bag of correlated poses is a relatively new area of research, but a wide variety of promising advantages are demonstrated in this method. Body poses encoded by silhouettes are considered to be robust to different clothing, appearance and illumination changes and it is the best way to detect motion. The extracted normalized silhouettes are used as input features for the Bag-of-Features (BoF) model. The proposed correlogram of body poses contains both statistical and temporal relationship information, which enables the algorithm to be capable of distinguishing actions with similar pose statistics but different temporal ordering.

Fig. 2. Detecting region of interest

In the interest point based action recognition method as shown in figure 2, each feature vector is a 3-D descriptor calculated around a detected interest point in an action sequence. In this method each feature vector is converted from the 2-D silhouette mask to a 1-D vector by scanning the mask from top-left to bottom-right pixel by pixel. Therefore, each frame at the time t in an action sequence is represented as a vector of binary elements, the length of which is

$$L = row * column \quad (1)$$

where "row" and "column" are dimensions of the normalized pose silhouette. Suppose the i^{th} action sequence consists of Si frames, then an action sequence can be represented as a matrix Xi with Si rows and L columns. Each row of the matrix stands for a single frame. Therefore, for a training set with n action sequences, the whole training dataset can be represented as

$$X = [X1;X2; \ldots ;Xn] \quad (2)$$

The total number of rows, which is also the total frame number in the training dataset, is

$$S = S1 + S2 + \ldots + Sn \quad (3)$$

Because features are in high-dimensional space, we first use PCA for dimensionality reduction. Hence, each frame Ft is projected into a lower dimension. Then, visual vocabulary can be constructed by clustering feature vectors obtained from all the training samples using the k-means. The size of the visual vocabulary is the number of the clusters k.

A color correlogram is a 3-D matrix where each element indicates the co-occurrence of two colors those are at a certain distance from each other. In action representation, each element in BoCP denotes the probabilistic co-occurrence of two body poses taking place at a certain time difference from each other. Since the poses are divided into k clusters, the dimensionality of the correlogram matrix at a fixed time offset Δt is $k * k$, where k represents the codeword number

$$\zeta(i, j; \delta t) = \Sigma W_{(i,t)} + W_{(i,t+\delta t)} \quad (4)$$

where δt specifies the time offset, W(i,t) is the frame Ft's visual word probability to cluster i. Figure(3) shows the correlogram construction.

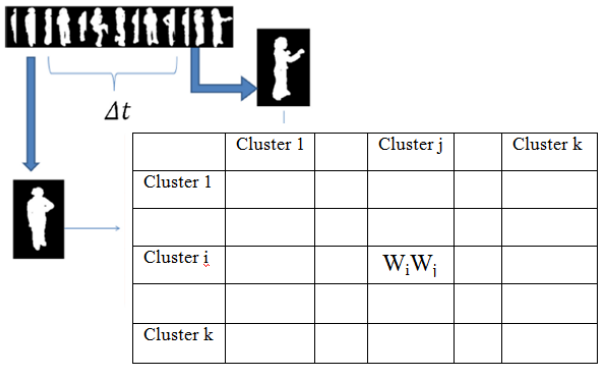

Fig. 3 Correlogram construction

B. *Dimensionality reduction*

Here we have done a comparative study of bag-of correlated method on three different feature vectors. Two linear and one nonlinear learning technique are used. The first one uses linear dimensionality reduction in order to find the underlying structure of the data. Both Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are used to learn a set of principal components (PCs) to characterize the data.

After pre-processing of the silhouette vectors, the dimension reduction process are proceeding to the training database contains the silhouette vectors with a high dimension. To extract the human activity silhouette features, the most popular feature extraction technique applied in the video-based HAR is Principal Component Analysis (PCA). PCA is the most widely-used and well-known of the standard multivariate methods. PCA consists on a transformation from a space of high dimension to another with more reduced dimension. If the inputs are highly correlated, there is redundant information. PCA decreases the amount of redundant information by decorrelating the input vectors. The correlated input vectors, with high dimension, can be represented in a lower dimension space and decorrelated. It is a linear projection method to reduce the number of

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

parameters. It transfers a set of correlate variables into a new set of uncorrelated variables.

Independent component analysis (ICA) is a computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals. ICA can be define by "Minimization of Mutual Information and Maximization of non-Gaussianity"

However dimensionality reduction techniques like PCA and ICA assume that the data essentially lies on a linear manifold. And it is very unlikely that they lie on a simple linear manifold. The main problem using PCA and ICA is that linear PCs cannot represent the non-linear nature of human motion.

The second method uses a non-linear dimensionality reduction technique. Specifically, spatio-temporal Isomap is applied to uncover the intrinsic non-linear geometry of the data, and it is captured through computing the geodesic manifold distances between all pairs of data points.

The crux of the Isomap algorithm is finding an efficient way to compute the true geodesic distance between observations, given only their Euclidean distances in the higher dimensional observation space. The idea is that Euclidean distance is approximately equal to the geodesic distance for close by points. For points which are far off the geodesic distance has to be computed by a series of hops. The Isomap algorithm as proposed in [11] consists of three main steps.

(1) Construct the neighbourhood graph G over all observation points.
(2) Compute shortest paths in the graph between using either the Floyd's or the Dijkstra's algorithm.
(3) Apply Multi Dimensional Scaling to the resulting geodesic distance matrix to find a d-dimensional embedding.

In PCA we try to preserve covariance. Here we try to do the same thing in a nonlinear way and our approach is to try to preserve inter point distance on the manifold. PCA and Isomap do this in an iterative fashion, trying increasing values for d and computing some sort of residual variance. Plotting this residual against d will allow us to find an infection point that indicates a good value for d.

IV. EXPERIMENTS AND RESULTS

In this section, the effectiveness of our motion recognition method is evaluated. The objective of this project is to recognize the categories of the human actions shown in the input silhouette files of the public test dataset. For the experimental purpose two different datasets are used namely Weizmann and IXMAS.

From Weizmann dataset only 4 actions are used for testing. They are Bend, Jump, Run and wave2 action. Each dataset in Weizmann contain set frames in the .tif format of having 180 x 144 dimensions. IXMAS action include checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, and picking up. Each action dataset contain series of silhouette of 390x291 dimensions. Images are .jpg format which can be stacked to recognize an action. This dataset is very challenging, because actors in the video sequences can freely choose their position and orientation. There are also significant appearance changes, intra-class variations, and self-occlusions.

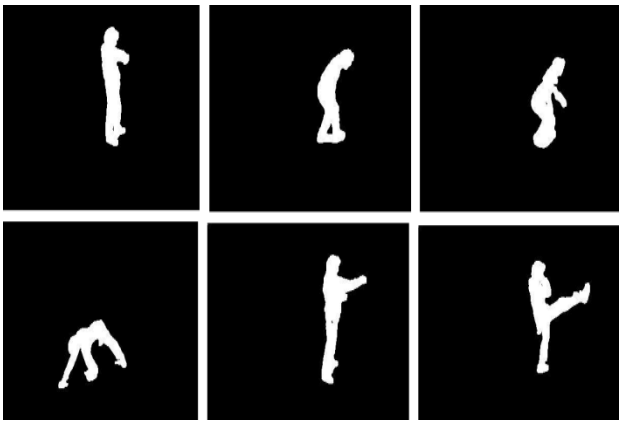

Fig. 4 Silhouettes of different actions:- check watch, Sit down, Stand Up, Pick Up, Punch and Kick

For the IXMAS dataset, only single view data is used for both training and testing and follow the widely adopted leave-one-actor-out testing strategy. Figure (4) shows some of the IXMAS silhouettes of check watch, sitting down, Standing up, picking up, punch, kick.

We experimented with three techniques, two linear techniques namely PCA and ICA, and one non-linear manifold learning technique called Isomap. The traditional PCA based approach that used Euclidean distance measure in the high dimensional space and the other was a non-linear technique. It is obvious that if data lies in a non-linear manifold, the PCA based approach would not model the data appropriately. To this end we used the non-linear Isomap technique that uses the Geodesic distance along the manifold. The evaluation on independent test set has shown that Isomap performs marginally better than the traditional PCA based technique.

We choose the following parameter settings: the bounding box of silhouettes is $30 * 20$ pixels and feature vectors are reduced to the dimension of 30 using different dimensionality reduction techniques. For each input frame a gray scale image is computed on which the blob region of interest is projected to a bounding box. Since the number of frames in each set is less than 100 the clustering index k for k-means should be less than 6. Correlogram matrix is created on these clustered images separately. It is found that there is visible difference in the Correlogram plotted for different action Both in PCA based and ISOMAP based methods. Intermediate results are stored on matrix named

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

"cormat" while reduced feature vector trained dataset is used for classification purposes. Multiclass SVM with 4 classes is trained first. Then testing is conducted on the same dataset.

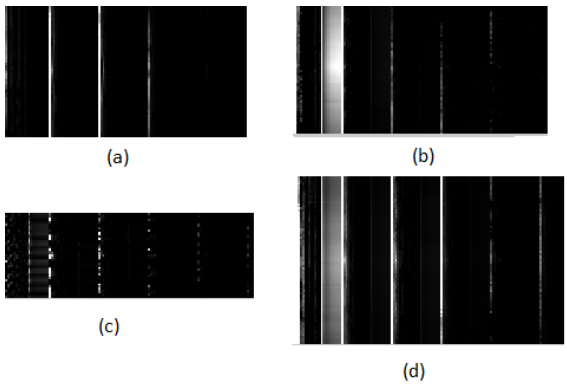

Fig. 5 A sample correlogram for Weizmann dataset usin ISOMAP (a) Bend (b) Jump (c) Run (d) Wave

On figure(5) there are correlogram matrices of different actions performed by the same person from the same set of action. It is even visually possible to distinguish the difference in texture between different actions' correlograms. We can observe that the correlogram matrices of the same action look much more similar than those of different actions, which makes correlogram a discriminative representation for human actions.

V. CONCLUSION

It is evident that human motion analysis plays a crucial role in advancing computer vision. This project has concluded with a simple but effective method for automatic person recognition from body silhouette with an extension of Correlogram based approaches called Bag of correlated posses (BoCP). BoCP is a temporally local feature descriptor. In the BoCP model, a unique way of considering temporal-structural correlations between consecutive human poses encoded more information than the traditional bag-of-features model.

In this work, our system showed promising performance and produced better results when using ISOMAP than using only low level features and simple dimensionality reduction methods like PCA and ICA. Even though ICA is proved to be a good method for action recognition our experiment shows that ICA is not apt for correlogram construction. For classification where accuracy is concerned ISOMAP is best but PCA can be used when speed is concerned. With more sophisticated feature descriptors and advanced dimensionality reduction methods, better performance can be achieved.

REFERENCES

[1] Aaron F. Bobick and James W. Davis "The Recognition of Human Movement Using Temporal Templates", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 23, NO. 3, MARCH 2001

[2] Lena Gorelick , Shechtman , Irani and Basri "Actions as Space-Time Shapes" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 29, NO. 12, DECEMBER 2007

[3] Xinghua Sun and Mingyu Chen Alexander Hauptmann "Action Recognition via Local Descriptors and Holistic Features", Computer Vision and Pattern Recognition (CVPR) Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference

[4] Di Wu, "Silhouette Analysis-Based Action Recognition Via Exploiting Human Poses" IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 23, NO. 2, FEBRUARY 2013

[5] Fan Zhu, Ling Shao and Mingxiu Lin "Multi-view action recognition using local similarity random forests and sensor fusion", ELSEVIER Pattern Recognition Letters 34 (2013) 20–24

[6] Ling Shao, Ling Ji, Yan Liu, Jianguo Zhang, "Human action segmentation and recognition via motion and shape analysis", ELSEVIER, Pattern Recognition Letters 33 (2012) 438–445

[7] M Xiantong Zhen, Ling Shao "Embedding Motion and Structure Features for Action Recognition", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY ,2013, Unpublished

[8] Sen Wang, Yi Yang, Zhigang Ma, Xue Li, Chaoyi Pang, Alexander G. Hauptmann "Action Recognition by Exploring Data Distribution and Feature Correlation", Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference

[9] Michael B. Holte, Bhaskar Chakraborty, Jordi Gonzàlez, and Thomas B. Moeslund,"A Local 3-D Motion Descriptor for Multi-View Human Action Recognition from 4-D Spatio-Temporal Interest Points" IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 6, NO. 5, SEPTEMBER 2012

[10] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vision Image Understand.*,vol. 104, nos. 2–3, pp. 249–257, 2006.

[11] http:// isomap.stanford.edu/