Action Recognition Using Bag of Correlated Poses And Extended Motion History Image

Shaija N.
Department of Computer Science
College of Engineering, Chengannur
Alappuzha, India

Priyadarsini S.
Department of Computer Science
College of Engineering, Chengannur
Alappuzha, India

***Abstract*— This paper is proposed to explore a scheme for human action recognition incorporating the benefits of local and global representations. The correlation between sequential poses in an action is considered in the exploration of human silhouettes for human action representation. To encode temporally local features of actions, bag of correlated poses is used. The property of visual word ambiguity is utilized with the adoption of the soft assignment strategy. The loss of structural information can be made good with extended motion history image. Principal Component Analysis (PCA) is often used for reducing the dimensionality of input feature space. In this paper, we use the feature space based on Independent Component Analysis (ICA) and which is more effective than the PCA representation for human action recognition. Experimental results prove the practicability of the complimentary properties of two descriptors and also show that the ICA approach produces more accurate recognition than the PCA approach.**

Keywords—Action recognition, bag of correlated poses (BoCP), extended motion history image, soft assignment.

I. INTRODUCTION

Action recognition has become important over the past years. There are many potential applications, including video surveillance, human-computer interaction, gesture recognition, and video retrieval. The general method for human action recognition is to extract human motion information directly from video sequences and to compare it with a human action database. The important problem for human action recognition is how to learn and classify human actions efficiently.

Human action recognition consists of three major steps, human detection and tracking, feature extraction, and training or classification. The first step is to detect and track a human figure in an image sequence. Background subtraction is a simple method for human detection and tracking. For each image sequence, a background subtraction algorithm and a simple correspondence procedure are firstly used for segmenting and tracking the moving silhouettes of a human figure. The second step is to extract human motion features from the each frame and to normalize the image size with the shape centroid. The third step is to apply ICA for reducing the dimensionality of input feature space and to recognize human actions using standard pattern classification techniques in the lower-dimensional eigen space. The general human actions are named as follows: Walk, Run, Skip, Hand waving and so on. The experimental motion patterns of many researchers are approximately 5 ~ 10 patterns. Human action recognition is a very difficult problem, because human actions have a lot of motion patterns and there are the few open databases of human actions.

In this paper, we focus on the representation of actions using human poses, i.e., silhouettes, as our primitive descriptive units. We proposed two representations, namely, BoCP and the extended-MHI for action recognition.

II. BAG OF CORRELATED POSES

Fig. 1 shows the flow chart of the construction process for the BoCPs representation.

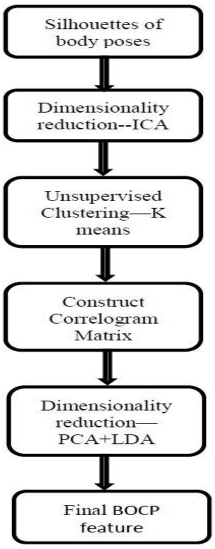
Silhouettes of body poses
Dimensionality reduction--ICA
Unsupervised Clustering—K means
Construct Correlogram Matrix
Dimensionality reduction—PCA+LDA
Final BOCP feature

Fig. 1. Flow chart of the BoCP model.

The initial idea of BoCP was presented in [1], where it was called correlogram of body 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. An action sequence is a series of pictorial frames. We treat each frame individually as an atomic input. The notion of body poses in our approach is represented by the silhouettes. A bounding box, i.e., the smallest rectangle containing the human figure, is applied to each frame of the silhouette sequence and then is normalized to a fixed size. The prepossessing steps reduce the original dimension and remove global scale and translation variations. The interpolation during the normalization process suppresses the noise as the morphological transformations (dilation and erosion) can isolate individual elements and join disparate elements in an image.

A. Codebook Creation

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[2] [3]. 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.

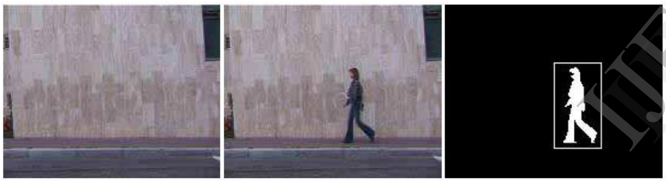

Fig. 2. Detecting region of interest

In the interest point based action recognition method [4] as shown in figure 2, each feature vector is a 3-D descriptor [5] 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

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

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

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

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

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

Because features are in high-dimensional space, we first use ICA 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 algorithm. The center of each cluster is defined as a codeword and the size of the visual vocabulary is the number of the clusters k.

B. Soft Assignment Scheme

The term "soft assignment" is commonly used in histogram comparisons. It describes techniques that identify a continuous value with a weighted combination of nearby bins, or "smooth" a histogram so that the count in one bin is spread to neighboring bins. The visual word vocabulary in the codebook model can be constructed in various ways, e.g., Gaussian mixture model, spectral clustering, and others. Typically, a vocabulary is constructed by applying k-means clustering. Here, an important assumption is that a discrete visual word is a characteristic representative of an image feature. The k-means algorithm minimizes the variance between the clusters and the data, placing clusters near the most frequently occurring features. However, the most frequent features are not necessarily the most discriminative and the continuous nature of visual appearance complicates selecting a representative visual word for an image feature. Assigning a feature to its single cluster gives rise to loss of information due to quantization errors, especially for features residing on boundaries of neighboring clusters. Thus, in our approach, we model our visual words by a kernel codebook to integrate the visual word ambiguity. Kernel density estimation is a robust alternative to histograms for estimating a probability density function. Because Gaussian kernel $K_\sigma = \exp(-1/2\ x2 ⁄ \sigma 2)$ assumes that the variation between an image feature and a codeword is described by a normal distribution with a smoothing parameter σ, we adopt this statistically viable kernel function.

C. Correlogram of Poses

The concept of correlogram was first introduced by Huang et al. [6] , where they used color correlogram for image indexing. A color correlogram [7] 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

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

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.

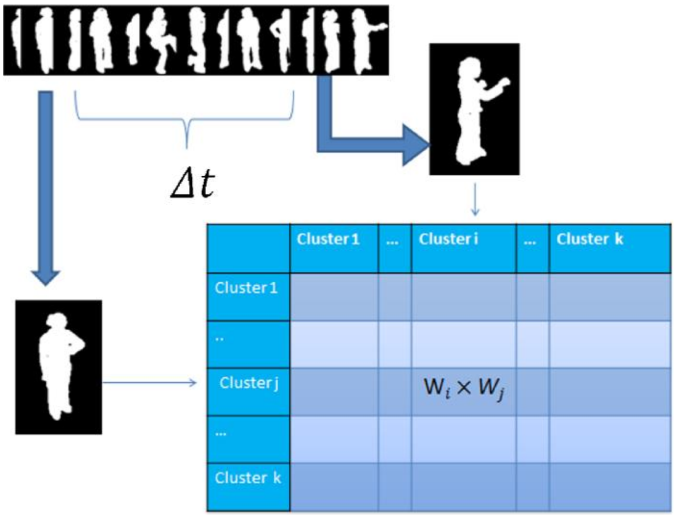

Fig. 3 Correlogram construction

D. Dimensionality Reduction

Both the original binary silhouette features and the final correlogram representations are in very high dimensionality. Due to the “curse of dimensionality,” it is impractical to use the original long feature vectors for classification. Therefore, the use of a dimensionality reduction method is necessary. There are two stages where dimensionality reduction is needed in our algorithm. The first stage is before feeding silhouette feature vectors into the k-means clustering process and the second stage is for the reduction of the final correlogram representations. We adopt the unsupervised PCA at the first stage and the combination of PCA and supervised LDA at the second stage. PCA seeks projection directions that maximize the variance of the data and LDA maps the features to make them more discriminative.

The accuracy of recognition is improved by using ICA [8] feature extraction. 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. 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"

At the first stage, a certain silhouette pose may appear in different action classes and the class label information is not very relevant. Therefore, an unsupervised method, i.e., PCA, is used. At the second stage, each correlogram matrix is at a high dimension (in a scale of 10^3-D); we first project the correlogram matrix into a lower dimension of 100-D using PCA and because each individual correlogram matrix belongs to one unique action class, we further reduce the dimension to the number of action class-1 using LDA.

III. Extended-MHI

As found by Sun et al. [1], local descriptors and holistic features emphasize different aspects of actions and share complimentary properties. Motivated by their finding, we fuse the above temporally local descriptor BoCP with an extension of the holistic descriptor: MHI by adding GEI and INV. These two additional holistic descriptors serve as the compensation for the loss of information due to sequentially overlapping frames lost in the original MHI representation. We deduce our approach by first introducing motion templates.

A. Motion Templates

MEI and MHIs proposed by Davis and Bobick [9] [10] are used to represent the motions of an object in video. Motion Energy Images (MEI) and Motion History Images (MHI) were introduced to capture motion information in images. They encode, respectively, where motion occurred, and the history of motion occurrences, in the image. Pixel values are therefore binary values (MEI) encoding motion occurrence at a pixel, or multiple-values (MHI) encoding how recently motion occurred at a pixel. Assume I(x, y, t) is an image sequence and let B(x, y, t) be a binary image sequence indicating regions of motion, which can be obtained from image differencing. The binary MEI E_τ (x, y, t) with the temporal extent of duration τ is defined as

$$E_\tau = \cup^{\tau-1} B(x, y, t-1) \tag{5}$$

The MHI Hτ (x, y, t) is used to represent how the motion image is moving, and is obtained with a simple replacement and decay operator as follows:

$$H_\tau(x, y, t) = \begin{cases} \tau, & \text{if } B(x, y, t) = 1 \\ \max(0, H\tau(x, y, t-1) - 1), & \text{otherwise.} \end{cases} \tag{6}$$

We further extend motion templates that include two more elements: GEI and INV.

GEI is to compensate for the nonmoving regions and the multiple motion instants regions of the action. INV is used to recover the loss of initial frames’ action information.

IV. Results And Discussions

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.

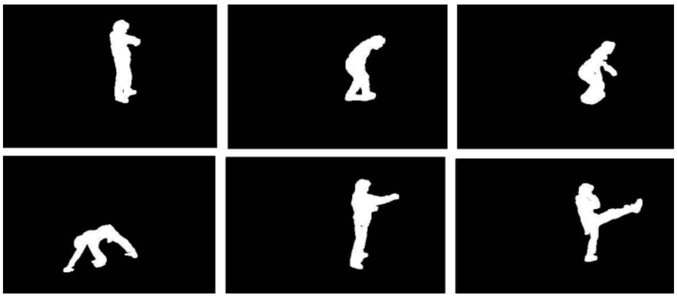

Fig. 4 Silhouettes of different actions:- check watch, Sit down, Stand Up, Pick Up, Punch and Kick

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 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. Each BoCP representation is then reduced to the dimension of the number of action class-1 using the combination of PCA (100-D) and LDA. Then, a unified framework is proposed to combine the two distinctive descriptors, BoCP and extended-MHI, by early fusion based on a very intuitive notion that the local descriptor (BoCP) and the holistic descriptor (extended-MHI) are complementary to each other. For final classification, the Gaussian kernel SVM classifier is adopted.

We also show the comparison results between PCA and ICA from the simple human action data. In order to select the number of independent components, we used PCA reduction before applying ICA. Recognition was done on the test sequences using a minimum Euclidean distance classifier and a Cosine similarity classifier that uses as similarity the angle between a test vector and a training one. The recognition rate was computed as the ratio of the number of samples classified correctly to the total number samples. Fig. 5 shows the recognition performance that was averaged over the recognition rate of the two motion features (silhouettes, frame difference) and the two classifiers (the minimum Euclidean distance, the Cosine similarity). It can be seen that the recognition rate of ICA was higher than that of PCA. The best recognition rate was about 75% for PCA and about 80% for ICA.

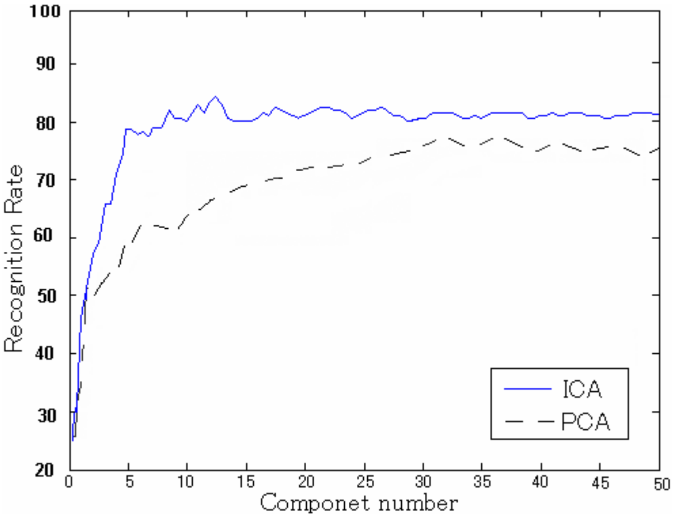

Fig. 5. Recognition performance of PCA and ICA.

V. Conclusions

In this paper, we proposed two representations, namely, BoCP and the extended-MHI for action recognition. 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. Also, we utilized a soft-assignment strategy to preserve the visual word ambiguity that was usually disregarded during the quantization process after *k*-means clustering. The extension of MHI compensated for information loss in the original approach and later we verified the conjecture that local and holistic features were complementary to each other. The ICA representation produces a better class separation than the PCA representation, since ICA can get the local motion features of human actions. With more sophisticated feature descriptors and advanced dimensionality reduction methods, we reckoned better performance.

ACKNOWLEDGMENT

The authors would like to thanks the anonymous reviewers for their constructive and useful comments, which helped in improving the presentation of our work.

REFERENCES

[1] X. Sun, M. Chen, and A. Hauptmann, “Action recognition via local descriptors and holistic features,” in Proc. IEEE Comput. Soc. Conf. CVPR Workshops, Jun. 2009, pp. 58–65.

[2] 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

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in Proc. 2nd Joint IEEE Int. Workshop Vis. Surveillance Performance Eval. Tracking Surveillance , Oct. 2005, pp. 65–72.

[4] I. Laptev and T. Lindeberg, “Space-time interest points,” in Proc. ICCV,2003, pp. 432–439.

[5] Ling Shao, Ling Ji, Yan Liu, Jianguo Zhang, “Human action segmentation and recognition via motion and shape analysis”, ELSEVIER.

[6] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, “Image indexing using color correlograms,” in Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog., Jun. 1997, pp. 762–768.

[7] L. Shao, D. Wu, and X. Chen, “Action recognition using correlogram of body poses and spectral regression,” in Proc. Int. Conf. Image Process., 2011, pp. 209–212.

[8] A. Hyvarinen, J. Karhunen, and E. Oja, IndependentComponent Analysis. Wiley, New York, 2001.

[9] J. Davis and A. Bobick, “The representation and recognition of human movement using temporal templates,” in Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog., Jun. 1997, pp. 928–934.

[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.