ACTION RECOGNITION USING CORRELOGRAM OF BODY POSES AND SPECTRAL REGRESSION

Ling Shao[1, 2], Di Wu[1], Xiuli Chen[1]
[1]Department of Electronic and Electrical Engineering, The University of Sheffield, UK
[2]Shenzhen Key Lab of Intelligent Media and Speech, PKU-HKUST Shenzhen Hong Kong Institution, Shenzhen, China

ABSTRACT

Human action recognition is an important topic in computer vision with its applications in robotics, video surveillance, human-computer interaction, user interface design, and multimedia video retrieval, etc. In this paper, we propose 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. To reduce the high dimensionality of the CBP representation, an efficient subspace learning technique called Spectral Regression Discriminant Analysis (SRDA) is explored. Experimental results on the challenging IXMAS dataset show that the proposed algorithm outperforms the state-of-the-art methods on action recognition.

Index Terms— Action Recognition, Histogram of Body Poses (HBP), Correlogram of Body Poses (CBP), Spectral Regression Discriminant Analysis (SRDA)

1. INTRODUCTION

Action recognition is a very active research topic in computer vision with a variety of applications, such as human-computer interaction, video indexing and retrieval, robotics, and visual surveillance, etc. Action representation is a critical component of action recognition and can be classified into global and local representations. Global representations (which sometimes are also called holistic representations) consider the action as a whole entity and do not require the detection and labeling of individual body parts. Background subtraction is usually applied to obtain a segmentation of the subject performing the action. Global representations retain the complete data of the action and are fundamentally more informative, but are sensitive to the segmentation accuracy and other complications such as partial occlusion and cluttering. Local representations treat the action as a collection of local features and the Bag-of-Features method is used to model the statistics of the features. Local representations are less informative, but usually more robust in challenging recording scenarios. A recent survey [1] summarizes different action representation methods.

We propose a new action representation based on the Correlogram of Body Poses (CBP), which has advantages of both global and local representations. The method takes silhouettes extracted using background subtraction as input information. Each silhouette encodes the body pose of a certain instant of the action and the statistics and temporal relationships of different poses are described by the Correlogram. Compared with many other existing methods, the new representation technique manifests several advantages:

1) Body poses are encoded by silhouettes, which are robust to different clothing, appearance and illumination changes;
2) We use raw data from normalized silhouettes as input for our action recognition, which saves the computation of feature description;
3) CBP 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;
4) The representation is also robust to unreliability of the segmentation, noisy labeling in training samples, and the speed of the action, because it is essentially a local representation temporally.

Because CBP has a high dimensionality, a recently proposed subspace learning method, Spectral Regression Discriminant Analysis (SRDA) [2], is used for efficient dimensionality reduction. SRDA aims to greatly decrease the computational complexity and is proven to be highly discriminative for action classification in our experiments.

In the remainder of the paper, we first introduce the new CBP action representation in Section 2. Section 3 briefly describes how SRDA is used in our algorithm. Experiments and results are presented in Section 4. Then, we conclude in Section 5.

2. ACTION REPRESENTATION

2.1 Silhouette representation

A human action can be viewed as a set of sequential silhouettes over time. Each silhouette records a pose of this action at a particular instant. There are a number of advantages for using silhouettes in human actions. Firstly, silhouettes are relatively easy to be obtained in many scenarios, especially when the camera is stationary which is widely available in both industrial and consumer surveillance systems. Given a static background, the

Fig. 1: Left: Illustration of the "cross arms" action; Right: A normalized silhouette.

background subtraction technique can be used to get reliable silhouettes. Secondly, silhouettes contain discriminative shape information, which is invariant to gender, body size, lighting condition, clothing, and appearance, etc.

A bounding box is applied on the 2D silhouettes and normalized to the same size. This simple step reduces the original dimension and removes global scale and translation variations. The interpolation during the normalization process suppresses the noise as well. The rectangular Region of Interest (ROI) mask serves as an image model in each action frame, making recognition invariant to body size as well as scale and translation variations resulting from perspective changes. Figure 1 shows a silhouette sequence of the action "cross arms" and an example of a normalized silhouette in a bounding box.

2.2 Correlogram of Body Poses

The extracted normalized silhouettes are used as input features for the Bag-of-Features (BoF) model. The difference to a common BoF model in action recognition is that body poses based on normalized silhouettes are used as features instead of local features extracted using spatio-temporal interest point detection and description. Due to the use of pose silhouettes, the local feature detection and description steps in a normal BoF method is much simplified. After the pre-processing in Section 2.1, the silhouettes contain the body pose information and have the same dimension. In interest point based action recognition, each feature vector is a descriptor calculated around a detected interest point in an action sequence. In our method, each feature vector is converted from the 2D silhouette mask to a 1D vector by scanning the mask from top-left to bottom-right pixel by pixel. Therefore, each frame in an action sequence is represented as a vector with binary elements, the length of which is E=row*column, where row and column are dimensions of the normalized pose silhouette. Suppose the i^{th} action sequence consists of N_i frames, then an action sequence can be represented as a matrix W_i of N_i*E dimension. Each row of the matrix stands for a single frame. Therefore, the whole training dataset can be represented as sample:

$$\mathrm{I} = [W_1; W_2; ...; W_i; ...; W_{total}] \tag{1}$$

The total number of rows, which is also the total frame number in the training dataset, is $N=N_1+N_2+\ldots N_i\ldots+N_{total}$. Then a visual vocabulary can be constructed by clustering the 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 therefore the number of the clusters. Each feature vector in I can be assigned to its closet codeword based on the Euclidean distance. Each action sequence from either the training set or the testing set can be represented as the probability distribution of the codewords in the form of histogram, which we call Histogram of Body Poses (HBP) [4].

HBP is a derivative of the Bag-of-Features model and only considers the probabilistic distribution of features but disregards the relationship among the features. Specifically, the temporal ordering of body poses is not taken into account, i.e. actions composed of the same body poses performed very differently sequentially will result in the same HBP.

The proposed Correlogram of Body Poses (CBP) representation attempts to take into consideration of both the statistical distribution and the temporal relationship of body poses. The concept of correlogram was first introduced by Huang et al. [5], where they used color correlograms for image indexing. A color correlogram is a three-dimensional matrix where each element indicates the co-occurrence of two colors which are at a certain distance from each other, which makes a correlogram much more informative than a histogram. Let I be an n*n image. The colors in I are quantized into m colors $c_1,\ldots,c_m$. The color correlogram of an image is defined as [5]:

$$\gamma_{c_i,c_j}^{(k)}(\mathcal{I}) \triangleq \Pr_{p_1 \in \mathcal{I}_{c_i}, p_2 \in \mathcal{I}}[p_2 \in \mathcal{I}_{c_j} \mid |p_1 - p_2| = k] \tag{2}$$

Inspired by color correlogram, CBP inherits the Bag-Of-Features advantages of HBP whilst incorporates the temporal pose relationship information. Each element in CBP denotes the co-occurrence of two body poses taking places at a certain time difference from each other. Therefore, the dimensionality of CBP matrix is K*K (K represents the number of codewords in k-means clustering). Each element in a CBP matrix can be defined as:

$$\begin{aligned} &C(cluster\_i, cluster\_j; \Delta t) \\ &= \sum_{frame=1}^{frameNo-\Delta t} W(cluster\_i, frame) * W(cluster\_j, frame + \Delta t); \end{aligned} \tag{3}$$

where Δt specifies the time offset which is a constant. The weight measurement W is defined as:

$$W(cluster\_i, frame) = \exp(-\|Center - Frame\|^2 / 2\sigma^2) \tag{4}$$

Here, *Frame* and *Center* denote the feature vector of a

Fig. 3: Flowchart of human action recognition.

particular frame and the centroid of a certain cluster, i.e. a visual word. Note also that here we use Mahalanobis distance instead of Euclidean distance. The Mahalanobis distance of a multivariate vector $x=(x_1,x_2,x_3,\ldots,x_N)^T$ from a group of vectors with the mean $\mu=(\mu_1, \mu_2, \mu_3,..., \mu_N)^T$ and the covariance matrix S is defined as

$$D_m(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{5}$$

In our experiments, we use the pooling covariance which achieves the best results.

There are two points that need to be mentioned here as the distinctive advantages of calculating the weights in CBP. Firstly, we preserve the maximum information by choosing the distance between an original frame and a cluster centroid as the weight. Secondly, the use of the extended Mahalanobis distance projects the nonlinear data to a kernel space which tends to suppress the problems of nonlinearity.

The CBP matrix can be obtained by assigning a number of different time offsets Δt. To some extent, the use of more time offsets would accordingly enhance the distinctiveness of the CBP representation but at the same time increase the dimensionality and complexity. In our implementation, four time offsets of 2, 4, 6, 8 are employed.

Figure 2 depicts examples of CBP. On the left side are CBP matrices of different actions performed by the same subject. It is even visually easy to distinguish the difference in texture between different actions. On the top right are CBP matrices of same actions performed by different persons. We can observe that the CBP matrices of the same action look much more similar than those of different actions, which makes CBP a discriminative representation for action sequences.

3. SPECTRAL REGRESSION DISCRIMINANT ANALYSIS

The dimensionality of the CBP representation is much higher than that of HBP. An efficient subspace learning technique is required to map CBP into a low-dimensional space. Linear Discriminant Analysis (LDA) has been a popular subspace learning method that preserves class separability. However, the computation of LDA involves dense matrix Eigen-decomposition, which can be computationally expensive in both time and memory. Specifically, LDA has $O(mnt+t^3)$ time complexity and requires $O(mn+mt+nt)$ memory, where m is the number of samples, n is the number of features, and $t=min(m,n)$. When both m and n are large, it becomes impractical to apply LDA.

In our experiments, we adopt a novel algorithm for discriminant analysis called Spectral Regression Discriminant Analysis (SRDA). Specifically, SRDA only needs to solve a set of regularized least squares problems and there is no eigenvector computation involved, which saves both time and memory tremendously. For details of SRDA, please refer to [2].

There are two stages where dimensionality reduction is needed in our algorithm, namely the feature vectors for pose silhouettes and the final CBP representation matrices. We use Principal Component Analysis (PCA) for the first stage and PCA+SRDA for the second stage. PCA seeks projection directions that maximize the variance of the data and SRDA maps the features to make them more discriminative. At the first stage, a certain 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 CBP matrix can only belong to one action class, and SRDA, which is a supervised method, should be more effective for the following classification.

4. EXPERIMENTS AND RESULTS

We evaluate our approach on a public action recognition dataset, the Inria Xmas Motion Acquisition Sequences (IXMAS)[1]. Each action is performed by 10 different subjects and sequences are recorded from different viewpoints with multiple cameras. This dataset is very challenging, because actors in the video sequences can freely choose position and orientation. There are also significant appearance changes, intra-class variations, and self-occlusions.

Figure 3 shows the flowchart of the training process of our algorithm. We use the traditional leave-one-out methodology for testing—9 persons are selected for training and the rest one for testing. The testing process is repeated for every possible combination and the results are averaged. The silhouette vectors are reduced to the dimension of 20 using PCA. During visual vocabulary construction, k=60 is used for k-means clustering, which results in 60 visual words. Since 4 time offsets are used in CBP, the dimensionality of a CBP representation is then 60x60x4=14400. Each CBP matrix is then reduced to the dimension of 10 using the combination of PCA and SRDA. For action classification, the K Nearest Neighbor (KNN) classifier is adopted.

Figures 4 and 5 depict the confusion matrices when using HBP or CBP as action representation, respectively.

[1] The data-set is available on the Perception website http://perception.inrialpes.fr in the "Data" section.

CBP gives much better results than HBP for all actions. It can be seen from the CBP confusion matrix that the recognition rates of "wave", "scratch head" and "check watch" appear to be much lower than others, partly due to their small changes in body poses. Table 1 shows the comparison to the state-of-the-art methods on the same dataset using a single camera view. Our algorithm based on the CBP representation outperforms the others significantly.

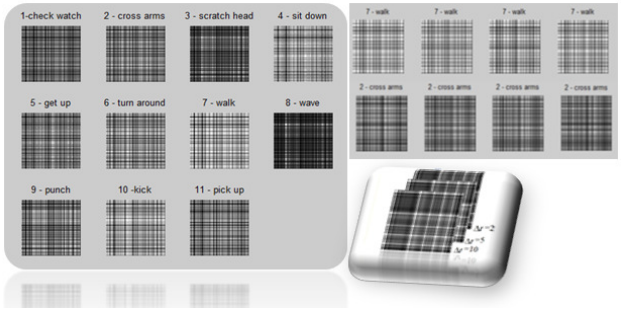

Fig. 2: Left: CBP matrices of different actions performed by the same person; Top Right: Same actions performed by different persons; Bottom Right: CBP matrices with different time offsets.

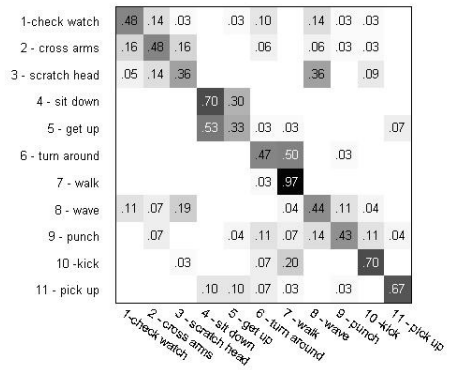

Fig. 4: Confusion matrix of action recognition using HBP.

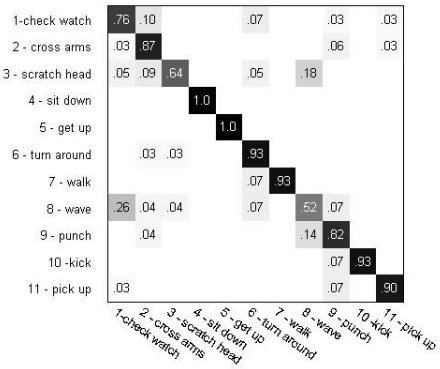

Fig. 5: Confusion matrix of action recognition using CBP.

6. CONCLUSION

In this paper, we propose a novel approach for action representation based on the Correlogram of Body Poses. Our approach is simple, efficient, and incorporates the advantages of both the global and local representations. In addition, we adopt a newly proposed subspace learning technique, called Spectral Regression Discriminant Analysis, for efficient regularized dimensionality reduction. Experimentally, we demonstrate that our algorithm achieves better performance than existing methods on the difficult IXMAS dataset.

Table 1: Recognition rates on the IXMAS dataset.

Authors	Recognition rate Single camera, Best result (%)	Method
Shao et al.	***85.6***	***CBP (Proposed)***
Lv and Nevatia [6]	80.6	PMK-NUP
Junejo et al. [7]	77.6	Temporal self-similarities
Weinland et al. [8]	81.3[1]	3D Exemplars
Yan et al. [9]	68.0	4D Feature Models
Weinland et al. [8]	63.5	2D Exemplars

1. Recognition rate is the result of 2 single camera combinations.

ACKNOWLEDGEMENT

The work described in this article was partially supported by the National Natural Science Foundation of China (Project no. 61005038).

REFERENCES

[1] D. Weinland, R. Ronfard, and E. Boyer, "A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition," Computer Vision and Image Understanding, 2010.

[2] D. Cai, X. He, and J. Han, "SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis," IEEE Transactions on Knowledge and Data Engineering, 2008.

[3] D. Weinland, "Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, 2006.

[4] L. Shao and X. Chen, "Histogram of Body Poses and Spectral Regression Discriminant Analysis for Human Action Categorization," British Machine Vision Conference, 2010.

[5] J. Huang, S. R. Kumar, M. Mitra, and W. J. Zhu, "Image indexing using color correlograms," US Patent 6,246,790, 2001.

[6] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[7] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "Cross-view action recognition from temporal self-similarities," European Conference on Computer Vision, 2008.

[8] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," International Conference on Computer Vision, 2007.

[9] P. Yan, S. M. Khan, and M. Shah, "Learning 4d action feature models for arbitrary view action recognition," IEEE Conference on Computer Vision and Pattern Recognition, 2008.