Multi Spatio-temporal Co-occurrence Measures for Human Action Classification

A. Q. Md Sabri[1,2], J. Boonaert[1], S. Lecoeuche[1] and E. Mouaddib[2]
[1]Mines Douai, Computer Science and Automatic Control Research Unit
764 Boulevard Lahure, Douai, France
[2]Université de Picardie Jules Verne, Modeling, Information and System Laboratory,
Faculty of Science, 33, rue Saint Leu, 80039 Amiens Cedex 1, France

Abstract

This paper deals with human action classification by utilizing spatio-temporal (ST) co-occurrences between labels of video-words that are stored within ST correlograms. Mutual information based clustering method is employed to reduce the size of the vocabulary created from local descriptors. Multiple characterizations for human actions in videos are extracted from the correlograms that are used for human action classification. These include a highly discriminative co-occurrence vector and a Haralick texture vector. The proposed method is implemented using a SVM classification technique. For evaluation purposes, the KTH and UCF Sports action recognition datasets, are used as they are the most well known and challenging datasets. The proposed method succeed in classifying different action classes, and improves the classification rate obtained by standard bag-of-video-words approach.

1 Introduction

Primary goal of our work is to improve the works of Savarese et al. [1] and our previous work [2], which utilize local based spatio-temporal co-occurrence technique. These works attempt to improve the standard bag-of-video-words (BoVW) model [3] used to characterize human actions. In recent years, mutual information (MI) has been successfully employed to compress the vocabulary of video-words for human action classification, [4], [5]. Motivated by this, we decide to incorporate MI based video-words selection into our proposed approach.

Savarese et al. [1] introduced vector-quantized representation of ST correlograms that describe co-occurrences of video-words within spatio-temporal neighborhoods. In [2], we highlighted that usage of a discriminative type of descriptor affects the overall ST texture which is represented by the ST correlogram. Thus, previously, we had chosen to use SURF based descriptor in replacement of the brightness gradient descriptors used by Savarese et al. [1]. This enhances the classification rate since the characterizations for actions contained in the different videos are better represented.

However, the main challenge in improving the works of Savarese et al. is that the information regarding the video-words labels that generate the co-occurrences is lost during vector quantization. Therefore, we propose to directly extract meaningful information from the ST correlogram without the usage of vector quantization.

Firstly, we propose a novel type of characterization for human actions by extracting a set of Haralick texture measures [6] from the ST correlograms. In image classification, texture measures are extracted from a distribution of pixel intensities in an image represented by a co-occurrence matrix [6]. In our work, similar texture measures are used to represent distinctive ST texture variations between different action classes. Secondly, another type of representation is obtained by reducing the dimensionality of the ST correlograms using PCA. Each of the ST correlogram is projected into a PCA subspace that reduces the number of its dimensions, preserving the important information. This in turn is used to characterize human actions.

Both the KTH [7] and the challenging UCF sports [8] datasets which are standard benchmarks for this area are used to evaluate our approach. Our proposed method succeed in classifying different action classes, and achieve near state of the arts performance.

2 Proposed Approach

Figure 1 and figure 2 depicts the global flow of our proposed approach for generating characterizations of human actions from an initial set of videos containing actions such as walking, jogging and hand-clapping.

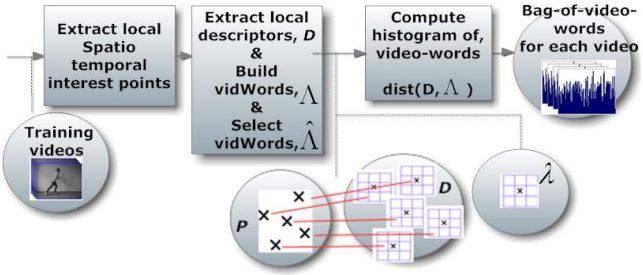

Figure 1: Bag-of-video-words approach

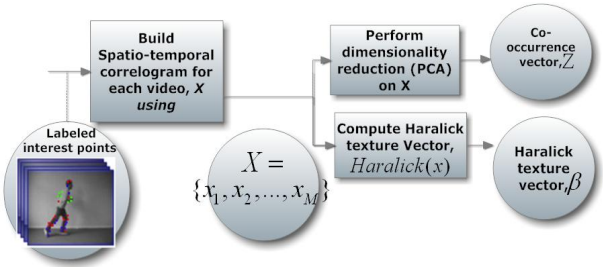

Figure 2: Spatio-temporal co-occurrence approach

2.1 Vocabulary of Video-Words

This section highlights the formation of video-words for the purpose of representing a video using the BoVW approach. BoVW approach starts with the detection of ST interest points (STIP) and its corresponding descriptors. STIP constitutes most salient areas in a

video indicating motion, and a descriptor is a patch surrounding the STIP that stores motion information.

Wang et al. [9] in their survey noted that the performance of local features are dataset dependent. In our work, we have chosen to use the Harris3D STIP detector with Histogram of Oriented Gradient (HOG) concatenated with Histogram of flow (HOF) [10] for both the KTH and UCF dataset. The implementation of these methods are available on the author's website[1], and default parameters are utilized in our experiments. We avoid dense sampling since, although it is effective in modeling cluttered scenes, it involves high computational cost as the number of STIP detected is often at least 10-20 times larger than sparse based STIP such as the Harris3D detector.

A video-word can then be considered as a representative of several similar patches. This is done by performing k-means clustering over the set of local descriptors. Video-words are then defined as the centers of the learned clusters. Thus, each patch in a video is mapped to a certain video-word through the clustering process and the video can be represented by the histogram of the video-words. This histogram is what we refer as BoVW representation of the video.

In brief, having n number of STIP detected, each interest point is mapped to its corresponding local descriptor,

$$P = \{p_1, p_2, ...p_n\}; D = \{d_1, d_2, ..., d_n\}$$
$$DESC : P \longmapsto D \quad (1)$$

Therefore, given a set of descriptors, D, extracted across a set of videos, K-means clustering is performed to partition the extracted descriptors into K number of clusters denoted by, Λ. The cluster centers are representative of the descriptors that are most prominent to describe and to discriminate these videos which contain different action classes such as running, walking and jogging.

$$KmeansWord : D \longmapsto \Lambda;$$
$$d_i \longmapsto \lambda_j;$$
$$j \in \{1, ..., K\}; i \in \{1, ..., n\}; K \ll n \quad (2)$$

A set of video-words is then defined as the centers of these clusters. This set is denoted by $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_K\}$. A label, l, for a video-word in this context refers to the index value associated with the video word (i.e. for λ_3 , l is then equal to 3) , whereby $l \leq K$.

2.2 Bag-of-video-words

Following the formation of a set of video-words Λ, and the extracted set of STIP, P, along with its corresponding set of local descriptors, D, each of the interest point is first labeled with the label l, of its nearest video-word. This is done by computing the Euclidean distance between the descriptor (associated with each interest point) D, and the stored video-words Λ.

$$Label : P \longmapsto P_{labeled}$$
$$p_i \longmapsto l_i \ ; \ l_i = \underset{j}{\operatorname{argmin}}(dist(d_i, \lambda_j)); \quad (3)$$
$$j \in \{1, ..., K\}; i \in \{1, ..., n\}$$

Frequency of video-words labels in a video is stored in a histogram used to characterize human action. This characterization will be jointly used with the the characterizations extracted using ST co-occurrence between the video-words labels.

[1] http://www.di.ens.fr/ laptev/download.html

2.3 Video-Words Selection

Often in BoVW characterization, a large number of video-words are extracted to generalize the dataset containing different actions. However, as our approach utilizes the co-occurrence between video-words labels, it is more desirable to have a subset of the initial vocabulary that preserves the initial classification rate to be used for implementing the ST co-occurrence technique.

To compress the size of initial vocabulary, we have adopted the approach presented by Liu et al. [4] for video-words selection based on mutual information (MI) between the video-words, Λ and the different action classes, Y.

In our implementation, utilizing the initial dictionary Λ obtained from K-means clustering, at each iteration, MI loss is computed for every pair of video-words, λ_1 and λ_2. The merging criteria is the pair that generates minimal MI loss. The merging process is continued until we reach the desired dictionary size, K^*. In our case, this is often to be between 20-40 percent of the original size of the vocabulary, and is selected based on the value of K^* that produces optimal classification rate. For more details on theoretical details behind the MI based video-words selection, we refer the reader to the works of Liu et al. [4]. After performing the video-words selection on Λ, we will obtain a reduced set of vocabulary, $\hat{\Lambda}$.

2.4 Construction of ST Correlogram and Extraction of ST Co-occurrence based characterizations

This section follows the notation and the works of Savarese et al [1] in generating the ST correlograms. From these correlograms we extract multiple ST co-occurrence based characterizations. These includes the Haralick texture vector and ST co-occurrence vector. Both types of characterization will be jointly used with the BoVW approach to characterize human actions. The set of vocabulary (i.e. video-words) referred in the following sections, unless specified, refers to the reduced set which is $\hat{\Lambda}$.

2.4.1 Spatio-temporal Correlograms

Following the formation of a reduced set of video-words $\hat{\Lambda}$, similar to the BoVW approach, the extracted set of STIP, P, along with its corresponding set of local descriptors, D, each of the interest point is first labeled with the label of its nearest video-word, $\hat{\lambda}$.

At this point, we have a distribution of labeled STIP. For each of the video from the set of videos containing different action classes, a local histogram $H(\Pi, p)$ is defined as a vector function that captures the number of interest points with the same label, l, within a spatio-temporal kernel Π, centered on p.

For each interest point location, p, a set of J kernels centered on p with different sizes is considered. This idea is taken from [1] and uses kernel type (rectangular volume) that extends between 2-40 pixels along the spatial dimension and 2-60 frames in the temporal domain. The r^{th} kernel of this set is denoted as Π_r. For concrete discussion on kernel construction and experiments related to kernel sizes, we refer readers to the original work by [1].

Table 1: Different Characterizations for Human Actions

Bag-of-video-words	Haralick Texture Measures	ST Co-occurrence vector
Histogram of occurrences of video-words labels	Energy Correlation Inertia Entropy Inverse Different Moment Sum Average Sum Variance Sum Entropy Difference Average Difference Variance Difference Entropy Info. Measure of Correlation1 Info. Measure of Correlation2	Dimensionally reduced ST correlogram vector

The average local histogram is defined as

$$\hat{H}(\Pi_r, l) = \sum_{p \in P_l} \frac{H(\Pi_r, p)}{|P_l|}; 1 \leq r \leq J; 1 \leq l \leq K^* \quad (4)$$

where P_l indicates the set of interest points with label l, and $|P_l|$ refers to its cardinality. A *correlogram*, x, for a particular video, $Vidx$, is built by concatenating in an array such local histograms for all combinations of labels and kernels.

$$x = \begin{bmatrix} \hat{H}(\Pi_1, 1) & \cdots & \hat{H}(\Pi_1, K^*) \\ \vdots & \ddots & \vdots \\ \hat{H}(\Pi_J, 1) & \cdots & \hat{H}(\Pi_J, K^*) \end{bmatrix} \quad (5)$$

2.4.2 Haralick Texture Measures

Haralick introduced different measures to extract texture information from 2D images. Details concerning the formula for the different measures can be referred to the original paper [6]. We consider the usage of Haralick texture measures based on the distribution of labeled video-words. Specifically, we wish to extract spatio-temporal texture information of different action classes, that is embedded within the distribution of the labeled video-words from each video containing a particular action. We utilized 13 different types of Haralick measures that are presented in Table 1.

Building blocks of a ST correlogram are local histograms of differing kernel sizes. Observing a particular kernel index, j, we can define $x_{\Pi j}$, which is essentially a co-occurrence matrix for all pairs of labels for a specific kernel size. More specifically, if we refer to row components of the previous definition from (5), we can define $x_{\Pi j}$ as

$$x_{\Pi j} = [\hat{H}(\Pi_j, 1), ..., \hat{H}(\Pi_j, K^*)]; j \leq J \quad (6)$$

We then proceed by extracting Haralick texture measures from each $x_{\Pi j}$. We can then define a mapping from a correlogram, x to its corresponding Haralick texture vector, β

$$\begin{aligned} Haralick : x &\longmapsto \beta \\ x_{\Pi j} &\mapsto \beta_m \\ j \in \{1, ..., J\}; m &\in \{1, ..., 13\} \end{aligned} \quad (7)$$

The Haralick texture vector, β, which is of the size $13 \times J$, is formed by concatenating each β_m extracted from each $x_{\Pi j}$ into a single vector.

This, along with other characterization types discussed can be used in combination (through vector concatenation) or separately to represent and characterize each video sequence.In our case, a naive implementation of the computation of the Haralick texture measures was performed, and therefore the maximal computational complexity as referred to [11] for a set of Haralick texture measures given a ST correlogram, is of the order $O(f^2)$, with f being the number of nonzero elements in each of the ST correlogram. However, as proposed by [11], there are methods to improve the efficiency of the computation of this measures.

2.4.3 Construction of ST Co-occurrence vector

In this section, we explain how we convert the ST correlogram into its corresponding ST co-occurrence vector. The goal of this approach is to avoid vector quantization that favors loss of information between the co-occurrences of video-words labels. At this point we have a set of ST correlograms extracted from a set of videos containing different actions, $X = \{x_1, x_2, ...x_M\}$, with M being the number of videos in a particular dataset.

A single ST correlogram is made up of $\hat{n} = K \times K \times J$ single-elements. Each correlogram, x_i is first transformed into a 1-d vector,$\hat{x_i}$, by concatenating all the rows of the ST correlogram into a single row.

Given a collection of these 1-D vectors, $\hat{X} = \{\hat{x_1}, \hat{x_2}, ..., \hat{x_M}\}$, we proceed by performing PCA to reduce the dimensionality of each $\hat{x_i}$. In our work, this is done by fixing the number of principal components, S, and by projecting each elements of $\hat{X}$ onto a new orthogonal basis, creating a new set of dimensionally reduced vectors, Z. This approach aims to reduce the number of dimensions while preserving the main information (i.e co-occurrences between video-words labels). The projected vectors, are smaller in dimension and contains essential video-words labels co-occurrences information that can be used to characterize human actions. The maximum value for S in our case is the number of training data used to compute the PCA values, and is equivalent to M.

The order of complexity to perform PCA on each of $\hat{x_i}$, is of order $O(\hat{n}D^2)$. The complexity is dependent on the size and dimension, D, of $\hat{x_i}$. Future works include the usage of kernel-PCA [12], in which PCA will be computed on the kernel values, removing this dependency.

$$\begin{aligned} PCA : \hat{X} &\longmapsto Z \\ x_i &\mapsto z_{\hat{i}} \\ i \in \{1, ..., \hat{n}\}; \hat{i} &\in \{1, ..., S\}; S \ll \hat{n} \end{aligned} \quad (8)$$

3 Experimental Results

In this section we will detail experimental results utilizing the different characterization methods discussed earlier for two distinct and highly challenging datasets, the KTH and UCF sports dataset.

3.1 KTH Dataset

The KTH dataset [7] contains 6 types of actions. There are 599 low-resolution (160 x 120) video files from a combination of 25 subjects, 6 actions and 4 scenarios taken indoor and outdoor. In our experiments, we followed the original experimental setup of the authors [7]. The dataset is divided into a test set (9 subjects), and training set (16 subjects). The initial size of the vocabulary, Λ used to create the BoVW histogram vector is 1000, in which we obtained 90.74 classification rate.

$\hat{\Lambda}$ of size 180 is created using the video-words selection method discussed earlier in which we preserved 88.89% classification rate. $\hat{\Lambda}$ is used to build the cor-

relogram from which Haralick texture vectors (Hara) and ST co-occurrence vectors (BOK) are extracted. These vectors are used to characterize different actions contained in training and testing set. For each of the training video, using one or the combination between the different type of vectors, along with its known action class, we train a supervised SVM classifier [12]. Table 2 denotes the results on the KTH dataset.

It is interesting to note that the usage of ST co-ocurrence vector alone (90.28%) is able to challenge the BoVW characterization (90.74%). Our approach propose the combination of all characterizations, in which we obtain a classification rate of 92.13% which is inline with the current state-of-the arts results reported on the KTH dataset, ([4], 93.43 %).

Table 2: Results for the KTH dataset

Action/ Charac.	BoVW(%)	BOK(%)	Hara(%)	Combined (%)
Box	100.00	100.00	88.89	100.00
Clap	97.22	100.00	58.33	97.22
Wave	91.67	97.22	77.78	91.67
Jog	83.33	80.56	30.56	91.67
Run	72.22	63.89	52.78	72.22
Wave	100.00	100.00	63.89	100.00
Avg.	90.74	90.28	62.04	92.13

4 UCF Sports Dataset

The UCF sports dataset is a collection of 150 broadcast sports videos and contains 10 different actions. It is a highly challenging dataset with large variations in terms of scenes and viewpoints. In our experiments, we adopted the experimental setup utilized by [5] that utilizes a 5-fold-cross-validation setup. The initial size of the vocabulary, $\mathbf{\Lambda}$ used to create the BoVW histogram vector is 1000, in which we obtained 73.17% classification rate. $\hat{\mathbf{\Lambda}}$ of size 300 is then created using the video-words selection method discussed earlier in which we preserved 60.05% classification rate. $\hat{\mathbf{\Lambda}}$ is used to build the correlogram from which Haralick texture vectors (Hara) and ST co-occurrence vectors (BOK) are extracted. These vectors are used individually or in combination, along with its known action class, to train a supervised SVM classifier [12].

Results in Table 3 demonstrates that we continue to see the trend in which our proposed approach of ST co-occurrence technique improves the BoVW characterization method. The combination of all the vectors achieve 75.68% classification rate. Typically, works that achieve >80% classification employs a more advanced and computationally expensive type of descriptors as reported by [5] that uses global features and [9] that uses dense sampling in their survey. We on the other hand utilizes sparse STIP and is able to obtain near 80% classification rate. This is more practical for a real-time application which is often the case in human action classification. Noticeable classification increase is achieved in "Dive"' and "Kick"' action classes, achieving between 7-20 percent jump classification increase. "Skate" continues to be a problematic action class, due to the fact of the sparse STIP used is not able detect distinctive region of interest. Using a more advanced type of STIP detector will solve this problem.

4.1 Conclusion and Future Works

In our work, we proposed the usage of a feature selection method to increase the efficiency in the creation of ST correlograms. We also proposed 2 types of ST co-occurrence based characterizations that when used in combination with the BoVW characterization approach, obtained a near state-of-the arts classification rate despite using a sparse STIP detector. Future works include identifying individual effect of the different types of Haralick texture measures on the performance of the proposed approach.

Table 3: Results for the UCF sports dataset

Action/ Charac.	BoVW(%)	BOK(%)	Hara(%)	Combined (%)
Dive	85.71	100.00	64.29	92.86
Golf	61.11	38.89	0.00	61.11
Kick	75.00	55.00	85.00	95.00
Lift	100.00	100.00	33.33	100.00
Ride	66.67	58.33	8.33	66.67
Run	76.92	53.85	15.38	76.92
Skate	33.33	0.00	33.33	25.00
Swg-bench	85.00	70.00	55.00	85.00
Swg-side	76.92	61.54	30.77	76.92
Walk	77.27	72.73	77.27	77.27
Average	73.79	61.03	40.27	75.68

References

[1] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *IEEE Workshop on Motion and video Computing, 2008. WMVC 2008*, 2008, pp. 1–8.

[2] A. Q. Md Sabri, J. Boonaert, S. Lecoeuche, and E. Mouaddib, "Human action classification using surf based spatio-temporal correlated descriptors," in *IEEE International Conference on Image Processing, 2012. ICIP 2012*, 2012.

[3] L. Fei-fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *In CVPR*, 2005, pp. 524–531.

[4] J. Liu and M. Shah, "Learning human actions via information maximization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, Jun. 2008, pp. 1–8.

[5] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes." in *ICCV*, D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, Eds. IEEE, 2011, pp. 707–714.

[6] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[7] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *In Proc. ICPR*, 2004, pp. 32–36.

[8] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach: a spatio-temporal maximum average correlation height filter for action recognition," in *In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[9] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *University of Central Florida, U.S.A*, 2009.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2008, pp. 1–8.

[11] E. Miyamoto and T. Merryman, "Fast calculation of haralick texture features," 2005.

[12] B. Scholkopf, A. Smola, and K.-R. Mller, "Kernel principal component analysis," in *ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING*. MIT Press, 1999, pp. 327–352.