Spatial-Temporal correlatons for unsupervised action classification

Silvio Savarese[1], Andrey DelPozo[2], Juan Carlos Niebles[3,4], Li Fei-Fei[3]

[1] Beckman Institute, University of Illinois at Urbana Champaign, USA

[2] Dept. of Computer Science, University of Illinois Urbana-Champaign, USA

[3] Dept. of Computer Science, Princeton University, USA

[4] Robotics and Intelligent Systems Group, Universidad del Norte, Colombia

Abstract

Spatial-temporal local motion features have shown promising results in complex human action classification. Most of the previous works [6],[16],[21] treat these spatial-temporal features as a bag of video words, omitting any long range, global information in either the spatial or temporal domain. Other ways of learning temporal signature of motion tend to impose a fixed trajectory of the features or parts of human body returned by tracking algorithms. This leaves little flexibility for the algorithm to learn the optimal temporal pattern describing these motions. In this paper, we propose the usage of spatial-temporal correlograms *to encode flexible long range temporal information into the spatial-temporal motion features. This results into a much richer description of human actions. We then apply an unsupervised generative model to learn different classes of human actions from these ST-correlograms. KTH dataset, one of the most challenging and popular human action dataset, is used for experimental evaluation. Our algorithm achieves the highest classification accuracy reported for this dataset under an unsupervised learning scheme.*

1. Introduction

Accurate human action classification is a fundamental problem in computer vision as well as an active field of research in recent years. However, it still remains a challenging task for computers to achieve robust action recognition due to cluttered background, camera motion, occlusion and geometric and photometric variances of the foreground person(s). A good example is shown in Fig.1. In this dataset, many different subjects perform the same action (e.g. walking, or hand waving) against different background (e.g. indoor, outdoor), recorded by a moving camera (e.g. zoom in and out).

Figure 1. Examples of the KTH database. In this dataset, many different subjects perform the same action (e.g. walking, or hand waving) against different background (e.g. indoor, outdoor), recorded by a moving camera (e.g. zoom in and out). It still remains a challenging task for computers to achieve robust action recognition under these conditions.

Recently, a number of works using spatial-temporal features of video clips have shown great promise in action recognition and classification. For example, Efros *et al.* [7] propose the use of motion descriptors based on optical flow to characterize motion in very low resolution video sequences. Their method relies on previously segmented and aligned clips of human actions. Blank *et al.* [3] represent actions by describing the spatial-temporal shape of silhouettes. Their approach relies on background subtraction to obtain a silhouette of the moving subject. Another line of work is that of Fanti *et al.* [8] and Song and Perona [23], which integrate multiple cues, such as velocities and positions of tracked feature points, to model actions with triangulated graphs. Our work is mainly inspired by Dollar *et al.* [6] and Niebles *et al.* [16]. In both of these works, fragments of spatial-temporal features are extracted based on salient changes of the appearance of the patch. These spatial-temporal features are local descriptors of a video clip, subtending a spatial-temporal cuboid. In [6], a simple but supervised framework is deployed to discriminate

human action classes via these local spatial-temporal descriptors. Specifically, the authors use an SVM classifier and achieve highly discriminative results. Recently, [14] propose to model human actions by segmenting the video in spatio-temporal super-pixels; [12] introduce a scheme for matching video sequences using intra-frame and inter-frame constrained local features; [15] introduce the tensor canonical correlation analysis for extracting correlation features across videos in the joint space-time domain.

Unlike the aforementioned contributions, Niebles *et al.* [16] have demonstrated the usage of spatial-temporal features in an unsupervised generative model. Local features are clustered into spatial-temporal codewords called *video words* and then the distributions of these video words are learnt by a probabilistic lantent semantic analysis (pLSA). Though this framework is unsupervised, it achieves competitive results compared to [6, 14, 12]. Interesting is also the work by [18] which have modified the pLSA model to include position information of the local features. Unlike [16], however, the method requires labeled and segmented video sequences as an input.

While these spatial-temporal local features have demonstrated their effectiveness in action classification, the works mentioned above fail to capture long range temporal patterns of action classes. Fig.2 shows an example where two classes share a similar distribution of video words resulting in an ambiguous representation for the unsupervised generative learning scheme. Our goal in this paper is to preserve the unsupervised nature of the generative models proposed by Niebles *et al.* [16], but incorporate the learning of discriminative temporal patterns across these local features. A number of works have proposed to learn temporal trajectories of motions (e.g. Yilmaz and Shah, [24]). But most of these papers rely on a tracking algorithm that might or might not provide reliable information. In addition, the temporal pattern that is learnt via these methods is imposed to be the 'global' temporal patterns of the entire action sequence. The coverage of the temporal window is pre-designed and inflexible. Ideally, one would like to learn these patterns at flexible temporal scales so that the algorithm is more robust to occlusion in the temporal dimension.

We introduce the idea of correlograms to capture the temporal co-occurrences patterns of the spatial-temporal features. This idea is initially proposed by Savarese *et al.* [19] for learning object classes in static images. The correlograms are used to capture the long-range spatial patterns beyond the local appearance of patches. We suggest that a similar idea can be extended in the time domain. In this paper, we propose to use *spatial-temporal correlatons, or ST-correlatons* to model the temporal correlation patterns among these spatial-temporally local codewords in different human action classes. We will give the formal definition of the ST-correlatons in Sec.2. We test our algorithm in the most challenging human action dataset: the KTH action dataset [20]. We obtain the best action classification results reported for this dataset under an unsupervised learning scheme.

The rest of the paper is organized in the following way. Sec.2 introduces the concept of correlogram and show how we can modify and extend it to the spatial-temporal domain. Sec.3 illustrates an action classification model based on these spatial-temporal correlograms. We show the experimental results and comparison with previous works in Sec.4 and conclude the paper in Sec.5.

2. Spatial-temporal correlograms

The original concept of correlogram goes back to Juletz [2] Haralick [17]. The idea was to describe the two-dimensional spatial dependence of gray scale values by a multi-dimensional structure called the co-occurrence matrix. Such matrix encodes the correlation of pairs of gray scale values as function of distance and angular variations and in turn provides powerful models for image classification. Huang *et al.* [11] extend this work and simplify the structure of the co-occurrence matrix by only retaining the correlation of pairs of intensity values (here the color values are considered as well) as function of distance. The shape context descriptor proposed by Belongie et al. [1] generalizes and complements these ideas for images with binary labels. Recently, Savarese *et al.* [19] have suggested to usage of correlograms for capturing the spatial arrangement of codewords. Furthermore, the authors achieve compact spatial modeling without loss of discrimination through the introduction of adaptive vector quantized correlograms, which they call *correlatons*. A correlaton is representative of a typical spatial co-occurance of a pair of codewords. Then, object models are obtained as histogram of codewords and correlatons. The former captures the appearance information of the object in terms of distribution of typical appearence elements (i.e. codewords); the latter captures the distribution of typical spatial relationships, (or, correlations between codewords - i.e., correlatons).

In this paper we follow the same philosophy of [19] and extend it to the spatial-temporal domain. Given a video-sequence, the action may be represented as a collection of spatial-temporal interest points. Each interest point has an associated label (membership). The label can be assigned by associating each interest point to the closest element of a vocabulary of representative spatial-temporal interest points which are called video words. Fig.2 shows examples of a spatial-temporal distribution of interest points where each of them is assigned to a video word (depicted by a color code). As Fig.2 shows, several interest points present the same label (i.e. same color code), hence creating the presence of spatial-temporal patterns of labels. Thus, the idea is

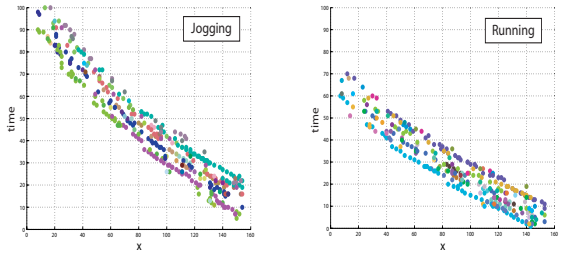

Figure 2. The action is visualized as distribution of video words. This distribution clearly forms patterns of visual words in time-space. The idea is to exploit correlograms to capture the spatial-temporal correlation of such patterns of labels. The graph depicts the top view of two different actions across time. The x axis represents the horizontal position on a x, y, t plot. Each color is associated with a different label. This figure is best viewed in colors

to exploit correlograms to capture the spatial-temporal correlation of such patterns of labels.

Similarly to [19], correlograms can be vector quantized, and we call the resulting quantized vectors *spatial-temporal correlatons* (or, *ST-correlatons)*. The set of ST-correlatons effectively captures the multi modal nature of spatial-temporal correlations of video words. Then, a video sequence can be characterized by the distribution of appearance elements (video words) and spatial-temporal relationships (i.e. ST-correlatons). This representation has the advantage of being very compact, reducing over fitting problems and yet being able to guarantee a high level of inter class discriminability as the experimental section shows.

2.1. Definitions

We follow the notation in [19]. Given a spatial-temporal distribution of spatial-temporal interest points and their corresponding label, we define a *local* histogram $\mathbf{H}(\Pi)$ as a vector function which captures the number of occurrences of each label i within a kernel Π (Fig.3). For each interest point location $\mathbf{p}$ we consider a set of K kernels centered on $\mathbf{p}$ and with different shapes and size (Fig.4). The r^{th} kernel of this set is denoted as Π_r. For each kernel we can compute the corresponding local histogram $\mathbf{H}(\Pi_r, \mathbf{p})$. Now, since we want to capture the correlation between group of spatial-temporal interest points sharing the same label, we define the *average* local histogram $\hat{\mathbf{H}}(\Pi_r, i)$ as the vector

$$\hat{\mathbf{H}}(\Pi_r, i) = \sum_{\mathbf{p} \in \{\mathbf{p}_i\}}^{|\mathbf{p}_i|} \frac{\mathbf{H}(\Pi_r, \mathbf{p})}{|\mathbf{p}_i|} \qquad (1)$$

where $\{\mathbf{p}_i\}$ is the set of interest points with label i and $|\mathbf{p}_i|$ is its cardinality. A *correlogram* is built by concatenating in a matrix structure such local histograms for all combinations of labels and kernels. If we fix a pair of labels i, j and we observe the co-occurences of i, j as function of the kernel type, we obtain a *correlogram element* $\mathbf{V}(i, j)$. Examples of correlogram elements are shown in Fig.5. Since correlograms are computed in a spatial-temporal domain we will denote them with *ST-correlograms*

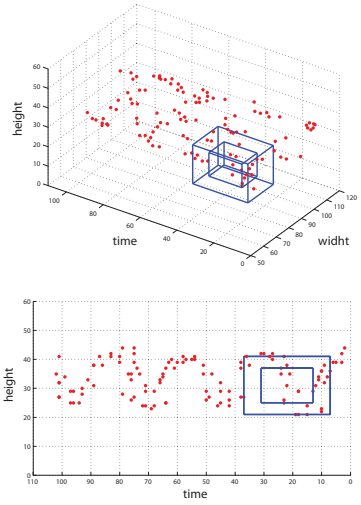

Figure 3. Examples of kernels and their location in the spatial-temporal pattern of codewords. The membership label (color) of each video word is not shown for clarity reasons.

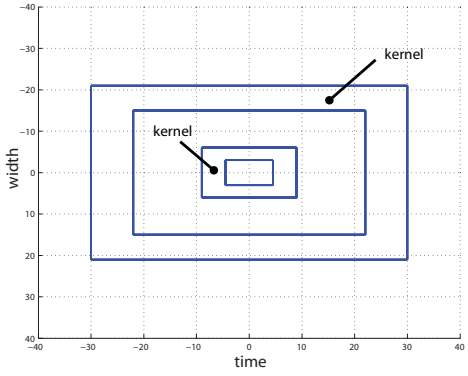

Figure 4. Two examples of kernel shape and size. The larger kernel has a maximum size of 60 frames along the temporal dimension and 40 pixels along the spatial dimension. The kernel "thickness" is about 10 frames times 5 pixels

2.2. Choice of the kernel

As opposed to [19] where only a single kernel with radial symmetry was used, here we build our correlograms with a set of kernels of various shapes. Fig.3 and Fig.4 show some examples. The motivation for this relies on the observation that different combinations of kernels of different shape and size may be more or less effective in capturing the underlying distributions of video words across categories. Our ex-

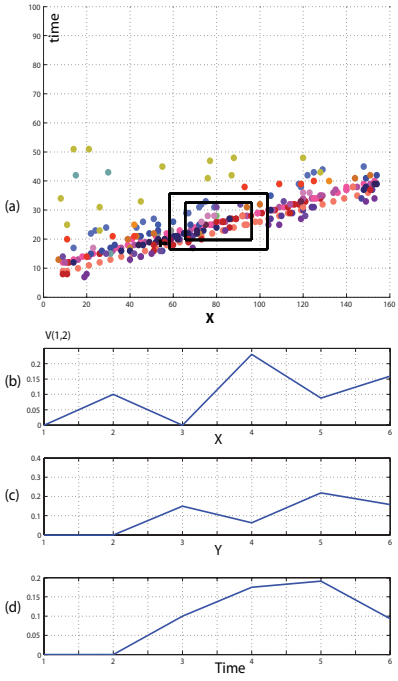

Figure 5. Examples of ST-correlogram element. Panel (a) shows a distribution of video words. A ST-correlogram element with respect a pair of video words labels (i.e. two color codes) is computed. The ST-correlogram element is shown in panels (b-d) and it expresses the co-occurrence of the pair of video words as function of the kernel shape and size. Panels (b), (c) and (d) express such co-occurrence as the kernel increases its size along the X, Y and T axis respectively.

periments show that this is indeed the case. We choose to use cubic-shaped kernels (as opposed to spheric ones) for reasons of computational efficiency (see in [19] for details).

3. Categorization based on spatial-temporal correlograms

In this section we describe how to use video words and ST-correlatons for modeling human actions classes. We first review the technique for extracting video words and ST-correlaton as well as for modeling human action classes based on distributions of video words. Then we review a unsupervised learning scheme based on the pLSA graphical model for building models of video sequences. Finally, we describe how to achieve categorization.

3.1. Representation by distribution of video words and ST-correlatons

The first step is to detect space-time interest points from the video sequence. We use a technique based on separable linear filters as proposed by Dollar *et al.* [6] and modified by Niebles *et al.* [16]. See relevant papers for details. The output of this algorithm is an interest point located in the video sequence and shaped as a space-time cube. For each spatial-temporal cube, we extract a descriptor formed by concatenating brightness gradients into a vector. The descriptors are clustered using a K-means algorithm. The resulting cluster centers form the vocabulary of video words.

Once the vocabulary of video words is computed, for each video sequence, spatial-temporal interest points are labeled with the closest video word in the vocabulary. Then, for each pair of video word labels we compute the corresponding ST-correlogram element by using a given set of kernels. We collect these vectors for all relevant video sequences in the training set and cluster them by using a K-means algorithm. The result is the set of representative ST-correlogram elements which we have called ST-correlatons. Notice that during the K-means clustering the membership of the ST-correlogram elements is lost (i.e. we do not know anymore which pair of video word have generated which)

3.2. Learning action classes via pLSA

Probabilistic topic models, especially probabilistic Lantent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) were initially proposed as a technique for semantic processing of the text documents [10, 4]. They are first introduced into the computer vision realm by Fei-Fei & Perona [9] and Sivic *et al.* [22] as a framework to learn scenes and object classes. More recently, Niebles et al. [16] demostrated the application of pLSA to learn human action classes from video sequences. Generally, the application of this model to vision tasks implied to adopt the "bag of keypoints" assumption, where the spatial-temporal arrangement of visual interest points is ignored. In this paper, we relax this assumption, by augmenting the pure "bag of keypoints" representation with the spatial-temporal relationships captured by the correlatons. Next, we review the statistical formulation of the pLSA model, following the presentation in [22] and [16] and extending it to our proposed method.

Suppose we have a training set of M video sequences, each containing observations from an augmented vocabulary that encodes appearance and spatial-temporal correlations. The size of the vocabulary is denoted as V. The corpus of training examples can be represented by a co-occurrence table $\bar{M}$ of size $V \times M$. The elements of $\bar{M}$ are denoted by $m(w_i, d_j)$, with $i = 1 \dots V$ and $j = 1 \dots M$, which count the number of occurrences of codeword w_i in video d_j. Note that the nature of w_i can be of appearance (video word) or spatial-temporal co-occurrences (correlaton), since our augmented vocabulary incorporates both types of information.

An intermediate latent variable z_k is also introduced, and associated with each occurrence of codeword w_i in a particular sequence d_j. We enforce the number of topics K to

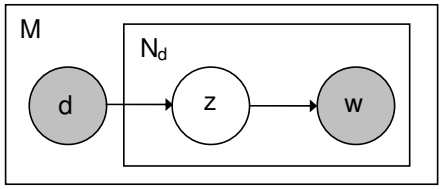

Figure 6. The *probabilistic Latent Semantic Analysis* (pLSA) graphical model. Nodes are random variables. Shaded ones are observed and unshaded ones are unobserved. The plates indicate repetitions. This figure is reproduced from [5].

be equal to the number of action classes to be discovered, so that each z_k represents one of the categories such as jogging, hand-waving, etc.

Following the relationships of the model variables, as shown in Figure 6, the joint probability can be written as:

$$P(d_j, w_i) = P(d_j)P(w_i|d_j) \tag{2}$$

The observed pairs (d_j, w_i) are assumed to be generated independently, and $P(w_i|d_j)$ can be obtained by marginalizing as follows:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k) \tag{3}$$

Here, $P(z_k|d_j)$ is the probability of action class z_k occurring in video d_j. Also, $P(w_i|z_k)$ represents the probability that codeword w_i occurs in a particular action category z_k. We can interpret this model as explaining the representation $P(w_i|d_j)$ of a particular video sequence as a linear combination of action specific codewords distributions $P(w_i|z_k)$ with weights $P(z_k|d_j)$.

Learning using the pLSA framework consists of estimating the model parameters $P(w_i|z_k)$, which are common to all examples, and $P(z_k|d_j)$, which are specific to each training point. The parameters can be learnt using EM, by maximizing the likelihood term:

$$\prod_{i=1}^{V}\prod_{j=1}^{M} P(w_i|d_j)^{m(w_i,d_j)} \tag{4}$$

where $P(w_i|d_j)$ is given by Equation 3.

3.3. Actions Categorization and Localization

The goal in recognition is to categorize novel sequences (d_{test}) given the learnt human action classes. The learning procedure has estimated the distribution of codewords for each particular action $P(w|z)$, and the task is to obtain the mixture coefficients $P(z_k|d_{test})$. This is again accomplished following a EM procedure. A categorization result can be obtained by choosing:

$$\text{Action Category} = \arg\max_k P(z_k|d_{test}) \tag{5}$$

Although the model is not particulary suitable for exact segmentation of the foreground data, an approximate localization of the performed action can be obtained by calculating the posteriors:

$$P(z_k|w_i, d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)} \tag{6}$$

By choosing the action class that maximizes $P(z_k|w_i, d_j)$ for each codeword w_i, we can effectively assign an action class to each codeword in the dictionary. When the codeword w_i corresponds to a video word instead of a correlaton, we can label its support region as containing evidence of the action class performed. In the case of w_i being a correlaton, the model captures likely co-occurrences instead of local region evidence. This becomes a subtle nuisance when defining its support region, and thus cannot be used for the purpose of localization in an straightforward manner.

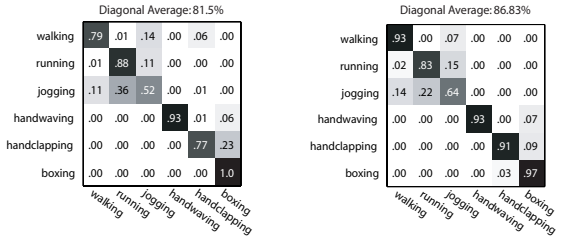

Figure 7. Confusion tables. The table on left shows the results by models based on distribution of video words. The table on right shows results by models based on the joint distribution of video words and ST-correlatons.

4. Experiments and results

In this section we assess the performance of our recognition scheme by reporting results for a set of experiments. We show that: i) models based on distribution of both video words and ST-correlatons systematically outperform models based on video words only; ii) different sizes of the vocabulary of video words and ST-correlatons lead to different performances, but models based on the combination video words and ST-correlatons achieve the best results consistently; iii) we explore performance dependency with respect to the different type and configuration of kernels for building the vocabulary of ST-correlatons. This indicates a possible design choice for an optimal set of kernels; iv) we report the results with models learnt on splitted sequences; this creates a more challenging test bed for models based on video words only (see details below) and shows that added spatial information helps boosting performance even more.

Our experiments are carried out on the KTH database[20]. This dataset is very challenging as it contains different configurations of video sequences, such as moving cameras, variable background and cluttered

scenes. The dataset features as many as six different types of human actions (see Fig.7). The database contains several instances of the same actions performed by 25 subjects. Different indoor and outdoors background environments are also used. See for example Fig.1.

The baseline for our experiments are the results presented in [16]. We use a similar learning strategy ([16] for details) based on the following key aspects: i) video words are built from two videos of each action from three subjects. ii) Recognition is carries through a leave-one-out (LOO) scheme. More specifically, at each round of the LOO iteration, a model from the videos of 24 subjects is learnt and tested on the videos of the remaining subject. The vocabulary of video words and ST-correlatons are learnt from videos of three subjects which are excluded from the testing beforehand.

We used a set of cubic-shaped kernels for building the vocabulary of ST-correlatons. The kernels size varies from 2 to 40 pixels along the two spatial dimensions, and from 2 to 60 frames along the temporal dimension. Notice that a video sequence is typically several hundreds of frames long, whereas a video words is only a few frames long, with an extension of a few pixels along the spatial dimensions. This suggests that correlatons may be useful to capture the co-occurrences of video words at different and intermediate spatial-temporal scales.

The results are summarized in Fig.7. The figure shows a side-to-side comparison of the two methods. The left panel presents results obtained by models in [16] with a vocabulary of video words of 500. Notice the large confusion between class *running* and *jogging* as well as between class *boxing* and *clapping*. These class share a similar distribution of video words resulting in an ambiguous distribution of topics for the pLSA model. Clearly, it is hard to learn models that guarantee the required discriminability without taking advantage of some additional degree of spatial-temporal information between visual words. The right-panel presents results obtained by using the joint distribution of video words and ST-correlatons. The size of vocabularies of video words and ST-correlatons is respectively 500 and 100. Performance is higher overall and the confusion between the *critical* classes is partially resolved. Fig. 8 shows examples where the joint distribution of video words and ST-correlaton resolves the ambiguity and allows the correct classification of these two instances. Table 1 compares performances of recent methods in the literature. Up to our knowledge, our method achieves the best results compared to current state-of-the art results using unsupervised learning, and reaches competitive results using supervised learning.

We assess performances with respect to different size of the vocabulary of both video words and ST-correlatons. The results are reported in Fig. 9 and Fig. 10. Fig. 9 compares the classification accurary for different models as function of the size of vocabulary of video words. Models based on joint video words and ST-correlatons systematically overperform those based on video words alone. Notice that models based on video words reaches a peak for a vocabulary size of 500 video words as reported in [16]. Models based on joint distributions reach a peak earlier $(200 - 300)$ and saturate from that point on. Fig. 10 compares the classification accuracy for models based on joint distributions as function of the size of the vocabulary of ST-correlatons. Notice that performances are fairly constant as the size of the vocabulary increases. This suggests that a vocabulary composed by small set of representative correlatons are sufficient to maintain the added accuracy. Finally, Fig. 11 shows performances as function of the kernel size.

Methods	**Recogn. accuracy** (%)	**Learning**
Our method	**86.83**	**unlabeled**
Niebles *et al.* [16]	81.50	unlabeled
Kim *et al.* [15]	95.33	labeled
Jiang *et al.* [12]	84.40	labeled
Dollár *et al.* [6]	81.17	labeled
Schuldt *et al.* [20]	71.72	labeled
Ke *et al.* [13]	62.96	labeled

Table 1. Comparison of different start-of-the-art methods. Our recognition accuracy is the highest on the KTH dataset using unsupervised learning.

Configuration	**50**	**100**	**200**	**500**
V. words + ST-Corr.	71.65	74.10	76.36	79.34
V. words only	68.61	70.20	69.36	69.74
ST-Correlatron only	55.05	63.54	67.62	72.00
Improv. over v. words	3.04	3.90	7.00	9.60

Table 2. Summary of the results obtained by various models (histogram of video words, ST-correlatons, joint video words and ST-correlatons). These results were computed using 300 video words and the training scheme in experiment 2.

We carry out a second experiment by splitting the video sequences of KTH database in two or four legs. This increases the size of the training set but at same time makes the classification task harder. For instance, the *walk* sequence consists of a long video of a person walking left-to-right, right-to-left twice. By splitting this sequence in four legs, we obtain shorter *walk* sequences each showing only one direction of walking. Learning models from the splitted sequences is harder due to the multi-modality of the two training sequences: each of them are assigned to the label *walk* regardless of the direction of motion. This difficulty is reflected to the overall performances as shown in Table 2. Performances of models based on video words drop to as low as 69%. However, models based on joint video words and ST-correlatons achieve an improvement of almost 10%.

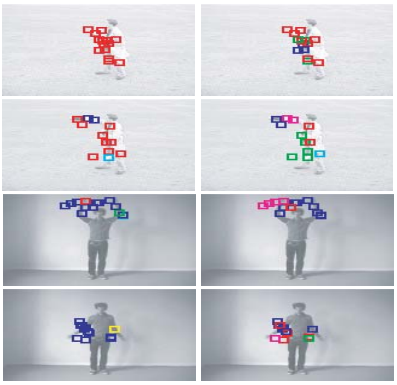

Figure 8. Examples of corrected mistakes. The right panel shows some instances where the appearance-only model misbehaves. Each color labels the action category. In this case, red for boxing and blue for handwaving. The left panel demonstrates that the join model increases the population of correct labels.

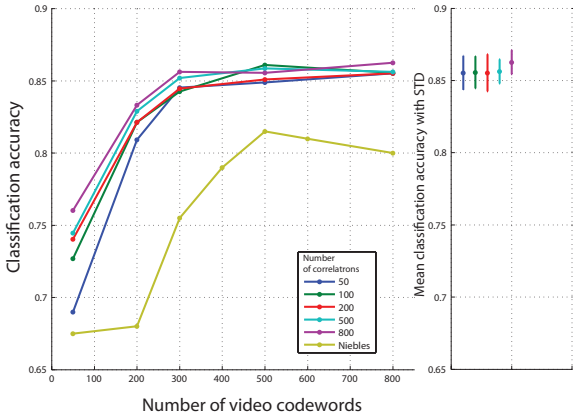

Figure 9. Performance as function of the size of the vocabulary of video words. The plots shows the improvement of our method over state-of-the-art results

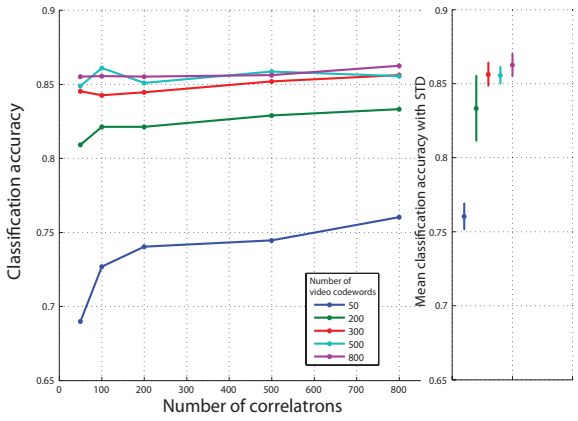

Figure 10. Performance as function of the size of the vocabulary of ST-correlatons.

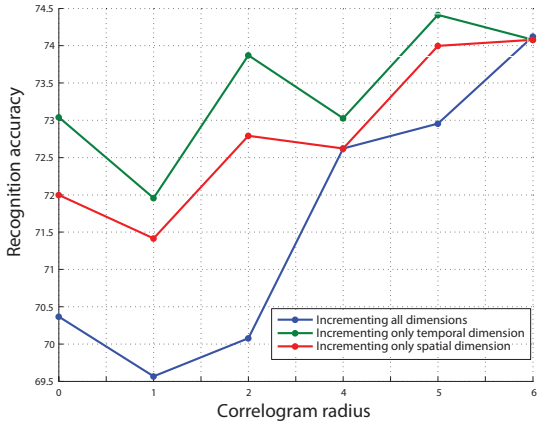

Figure 11. Performances as function of the size of the kernels. As the maximum size of the kernel is reduced performances drop significantly. The red plot shows performance drop when the maximum spatial dimension of the kernels is reduced from 20 pixels to 2 pixels; The green plot shows performance drop when the maximum temporal dimension of the kernels is reduced from 30 frames to 2 frames. The blue plot shows performance drop when the maximum of all dimensions are reduced.

5. Conclusion

Taking the simple idea of capturing temporal patterns for local motion features, our paper demonstrates that the usage of spatial-temporal correlograms proves to be highly effective in classifying complex human motions. We propose a framework for learning action class models from such correlogram features. Our results are very encouraging. We report the highest classification accuracy on the challenging KTH dataset compared to all previous methods, under unsupervised learning. We attribute the success to the flexibly learnt temporal signature of these motions. Much still needs to be done to further improve this work. For example, we would like to explore the possibility of adding additional global spatial and shape information into the model. Finding the optimal size/shape of kernels might be another direction of research. Finally, handling occlusions and multiple actions preformed at same time is still a fertile play ground for research. This effort should be coupled with the one of having available datasets comprising a more challenging kind of temporal/spatial variability.

6. Acknowledgments

This work is supported by a NSF grant 0413312.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[2] B.Julesz. Visual pattern recognition. 1962.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 2, pages 1395–1402. IEEE Computer Society, 2005.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, 2005.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 2, pages 726–733. IEEE Computer Society, 2003.

[8] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 1, pages 1166–1173. IEEE Computer Society, 2005.

[9] L. Fei-Fei and P. Perona. A Bayesian hierarchy model for learning natural scene categories. *CVPR*, 2005.

[10] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, August 1999.

[11] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *In Proc. IEEE Comp. Soc. Conf. Comp. Vis. and Patt. Rec., pages 762–768, 1997.*, 1997.

[12] H. Jiang, M. Drew, and Z. Li. Successive convex matching for action detection. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[13] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 166–173. IEEE Computer Society, 2005.

[14] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *7th Int. Workshop on Visual Surveillance*, 2007.

[15] T. K. Kim, S. F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[16] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *Proceedings of British Machine Vision Conference 2006*, volume 3, pages 1249–1258, September 2006.

[17] R.M.Haralick. Statistical and structural approaches to texture. In *Proc. of IEEE, 67(5):786-804*, 1979.

[18] T. K. S.-F. Wong and R. Cipolla. Learning motion categories using both semantic and structural information. In *In Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[19] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2033–2040, 2006.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, pages 32–36, 2004.

[21] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 405–412. IEEE Computer Society, 2005.

[22] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 370 – 377. IEEE Computer Society, October 2005.

[23] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(25):1–14, 2003.

[24] A. Yilmaz and M. Shah. Action sketch: A novel action representation. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 984 – 989. IEEE Computer Society, 2005.