Information Theoretic Key Frame Selection for Action Recognition

Zhipeng Zhao, Ahmed Elgammal
Computer Science Department, Rutgers University
110 Frelinghuysen Road, Piscataway, NJ 08854-8019, U.S.A
{zhipeng,elgammal}@cs.rutgers.edu

Abstract

This paper presents an approach for human action recognition by finding the discriminative key frames from a video sequence and representing them with the distribution of local motion features and their spatiotemporal arrangements. In this approach, the key frames of the video sequence are selected by their discriminative power and represented by the local motion features detected in them and integrated from their temporal neighbors. In the key frame's representation, the spatial arrangements of the motion features are captured in a hierarchical spatial pyramid structure. By using frame by frame voting for the recognition, experiments have demonstrated improved performances over most of the other known methods on the popular benchmark data sets.

1 Introduction

Recognizing human action from image sequences is an appealing yet challenging problem in computer vision with many applications including motion capture, human-computer interaction, environment control, and security surveillance. In this paper, we focus on recognizing the activities of a person in an image sequence from local motion features and their spatiotemporal arrangements.

Our approach is motivated by the recent success of "bag-of-words" model for general object recognition in computer vision[21, 14]. This representation, which is adapted from the text retrieval literature, models the object by the distribution of words from a fixed visual code book, which is usually obtained by vector quantization of local image visual features. However, this method discards the spatial and the temporal relations among these visual features, which could be helpful in object recognition. Addressing this problem, our approach uses a hierarchical representation for the key frames of a given video sequence to integrate information from both the spatial and the temporal domains. We first apply a spatiotemporal feature detector to the video sequence and obtain the local motion features. Then we generate a visual word code book by quantization of the local motion features and assign a word label to each of them. Next we select key frames of the video sequence by their discriminative power. Then, for each key frame, we integrate the visual words from its nearby frames, divide the key frame spatially into finer subdivisions and compute in each cell the histograms of the visual words detected in this key frame and its temporal neighbors. Finally, we concatenate the histograms from all cells and use

BMVC 2008 doi:10.5244/C.22.109

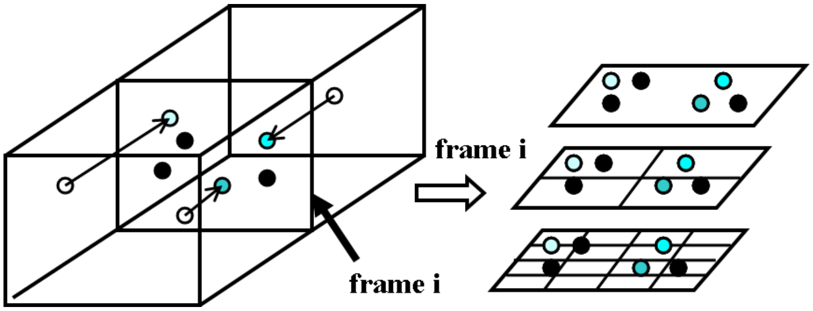


Figure 1: The representation for a key frame i is built from the motion features detected in it and integrated from the nearby frames. The closer a features is to the frame i, the higher weight it is assigned to, represented by a darker circle. Then a spatial pyramid (e.g. $L = 2$) is applied to model the spatial arrangements among the features.

it as the representation for this key frame. The process for representing a key frame i is illustrated in Figure 1. The contribution of our work lies in integrating the local motion features temporally into distinctive key frames, which can be used for efficient recognition. Besides the appearance information contained in the local motion features, our representation for the key frames also captures both the spatial and the temporal relations among the features, which leads to better performance than the popular “bag-of-words” approach.

2 Related Work

Extensive research has been done in recognizing human activities. The approaches can be broadly categorized as model based, spatiotemporal template based and “bag-of-words” based. Model based approaches for action recognition depend on locating and tracking body limbs in order to recognize the activity. That requires a model of the body, whether a 3D model or a 2D view-based model. We refer the reader to excellent surveys covering this topic, such as [2, 7, 12]. However, for the task of activity recognition, tracking the limbs is not necessary. That motivates research on obtaining spatiotemporal descriptors directly from the motion to recognize the action without limb tracking. One of the earliest work on spatiotemporal descriptor was carried out by Polana and Nelson[15]. In Bobick and Davis’s work[3], Motion-Energy-Image and Motion-History-Image are introduced as templates for different motion recognition. Efros et al. [6] also proposed a spatiotemporal descriptor based on global optical flow measurements. Spatiotemporal template approaches are holistic approaches where global descriptors are used with no local features extracted.

In contrast, “bag-of-words” based approaches detect local salient descriptors as visual words, which are then used to recognize the activity. “bag-of-words” has been used successfully for object categorization[21, 14]. Inspired by text categorization, it represents an object as a histogram of local features. Recently, “bag-of-words” methods have been used in action recognition[18, 4, 19]. However, these approaches lack the relations between the features in the spatial and the temporal domains which are helpful for recognition. There are many recent research on extending "bag-of-words" to add the spatial relation in the context of object categorization [17, 1, 11, 8, 10]. In particular, pyramid match kernel [8, 10] used the weighted multi-resolution histogram intersection as a kernel function for classification with sets of image features.

The approach we propose here tries to simultaneously model the spatial and the temporal relations of the local motion features for the key frames. The temporal information are captured by integrating the local motion features from the key frame's temporal neighbors and the spatial information are captured by using a hierarchical spatial pyramid in the representation.

Other related directions and extensions for "bag-of-words" in the context of action recognition include [9, 20, 22]. In [9]'s work, the spatial orientation information were captured in the local features. In [20, 22], latent semantic model was applied to discover the action types as topics in the hidden layer between the visual features and the video sequence. Savarese etc. [16] used correlograms to capture the spatial and temporal relations of local features. Different from them, our approach selects and models the key video frames and simultaneously integrates the spatiotemporal relation among visual features with their appearance information.

Feature selection has been used in object class recognition as a preprocess to build a better model. Different criteria, such as likelihood ratio and mutual information in [5], and different methods, such as combinatorial and statistical methods in [23], have been used for feature selection. Different from them, our approach selects the key frames by their entropy based discriminative power.

3 Spatiotemporal Representation for the Key Frame

3.1 Feature extraction

We use the feature extractor from Dollar[4], which has been proven successful in [4, 13, 22], for the detection and representation of the local motion features. In this method, the motion features are detected by applying separable linear filters to the video sequences. These local features are represented by the intensity gradients of a cuboid of spatiotemporally windowed data surrounding the detected interest point. To build the code book, we perform k-means from a random subset of motion features from the training data.

3.2 Key frames selection by their discriminative power

Intuitively, not all frames from a video sequence are equally important. We only need a few informative frames that characterize the action for recognition. The reasons are:

1) Some video frames are irrelevant to the underlying activity, e.g. the frames with no action in them. They could be nuisance for the recognition.

2) We can greatly speed up the recognition process if we only use the informative key frames without losing important information..

The feature exactor from Dollar[4] can detect the local informative motion features for each frame. They are encoded by the visual words $v_1, \ldots, v_K$, obtained from the clustering, where K is the vocabulary size. We can measure the discriminative power of each visual word. Entropy is a suitable measure for the discriminative power of a given visual word since it measures the uncertainty or the randomness of such a word. Given the set of activities $A_1, \ldots, A_N$, we can compute $P(A_i|v_j)$, the conditional probability of each action given visual word v_j, from the training data. The conditional entropy given the visual word v_j can then be computed as

$$E(Action|v_j) = \sum_{i=1}^{N} -P(A_i|v_j) log P(A_i|v_j) \tag{1}$$

The higher the entropy, the more uniformly distributed the activities are given the visual word, therefore, the less discriminative at the visual word. The lower the entropy is, the more discriminative the visual word is. We can use the conditional entropy of the visual words to measure the discriminative power of a given frame F. To do that, we use a function $g(\cdot)$ which is defined as:

$$g(F) = \sum_{j=1}^{K_F} \frac{1}{E(Action|v_j)} \tag{2}$$

where K_F is the number of the visual words in frame F. The higher the score of $g(F)$, the more discriminative the frame F is.

We selected the top p% most discriminative frames for recognition. p is set to 25 in the experiment, which is an empirical chosen number. These top discriminative frames are called the key frames.

3.3 Temporal integration of the motion features

Since a frame is correlated to its temporal neighbors, we build its representation from the motion features detected in it and its neighbor frames, weighed by the features' temporal distance to this frame. Intuitively, the further the distance is, the less weight it should be assigned to. Therefore, for a key frame i, the weights assigned to the motion features from a frame j are:

$$Weight(i, j) = e^{-\frac{dist(i,j)}{\sigma^2}} \tag{3}$$

where $dist(i, j)$ is the first norm distance between frame i and j and σ is the bandwidth for a smooth weight. The temporal relations of the features to frame i are captured by the different weights. The weights are 1 for the motion features detected at frame i and are close to 0 for those from the distant frames. Therefore, only the motion features from the nearby frames contribute significantly to the integration.

3.4 Spatial representation for the frame

With all the temporally weighted motion features for a given key frame, we need to find a representation to model the spatial relations of these features in a way that it is suitable for measuring the similarity between the key frames.

Let X and Y be two sets of motion features from two key frames respectively. Inspired by [10], we represent the key frame in a spatial pyramid. For each level $l, l = 0, \ldots, L$, we divide the key frame along x and y dimensions into $2^{2\times l}$ subdivisions. Intuitively, we measure the distance between X and Y as the sum of the distances between the corresponding cells of all levels from X and Y. Each cell can be described as the histogram of the weighted motion features in it and the distances between them are measured by Chi-square distance. So the distance between X and Y is formulated as:

$$dist(X,Y) = \sum_{l=0}^{L} \sum_{i=1}^{2^{2\times l}} \chi^2(H_X^l(i), H_Y^l(i)) \tag{4}$$

where $H_X^l(i)$ is the histogram of the weighted motion features from the ith cell in level l from X and $\chi^2(\cdot,\cdot)$ is the Chi-square distance. This is similar to representing X and Y by concatenating their histogram representations from all cells in all levels into a long histogram respectively and measuring their distance. Therefore, we can use these concatenated histograms as the representations for the key frames.

Since different information are captured at various levels of the pyramid, different weights should be assigned to each of them. At finer resolution, the correspondence between two sets are captured more accurately. Therefore, we penalize the similarity information gained at a coarser level and give more weights to the similarity measured by the histogram distance at a finer resolution. The weight we assign at level l is: $weight(l) = 1/2^{L-l}$ for $l = 0,\ldots,L$. The weighted distance between X and Y is:

$$dist(X,Y) = \sum_{l=0}^{L} \frac{1}{2^{L-l}} \times \sum_{i=1}^{2^{2\times l}} \chi^2(H_X^l(i), H_Y^l(i)) \tag{5}$$

Because Chi-square distance satisfies

$$c\chi^2(a,b) = \chi^2(ca,cb) \tag{6}$$

where c is a scalar, we can directly embed the weight to the histogram representation. Putting everything together, our representation for a key frame is the concatenated weighted histogram from all cells in all levels of the pyramid. In our representation, the temporal relations are modeled as the different weights assigned to the motion features and the spatial relations are captured in the spatial pyramid structure. With the motion features as the visual words, our representation simultaneously integrate appearance, spatial and temporal information.

Our representation for a given key frame is a straightforward extension of the popular "bag-of-word". In each subdivision, all the local motion features are modeled as "bag-of-words". When $L = 0$ and $\sigma = 0$, it reduces to the standard "bag-of-word" representation for the key frame. For better computation efficiency, we normalize the vector by the total weight of all its elements.

The complexity of the key frame representation is linear with the size of motion words vocabulary. For L level and K motion words, the dimensionality of the resulting representation is $K\sum_{l=0}^{L} 4^l = K\frac{1}{3}(4^{L+1} - 1)$.

3.5 Recognition algorithm

In the training phrase, for each training video sequence, we select the key frames, represent them using our spatiotemporal representation and label them with the underlying action. In the testing phrase, for each testing video sequence, we also select the key frames

Dataset	Facial Expression	Hand Gesture	KTH
No. of classes	6	9	6
No. of subjects	2	2	25
No. of trials per subject	8	10	1
No. of conditions	2	5	4
Total No. of Samples	192	900	593

Table 1: Details of the data sets used in our experiments.

and represent them with our spatiotemporal representation. Then these key frames from the testing video sequence can serve as weak classifiers. We use the nearest neighbor algorithm to label them with the closest key frames from the training data sets. Finally we combine these weak classifiers for the action recognition by employing a majority votes from them throughout the sequence.

4 Experiments

4.1 Data sets and experimental setting

We carried out our experiments in three data sets, namely facial expressions data set from Dollar et al.[4], hand gestures data set from Wong et al.[22] and KTH human action data set from Schuldt et al.[18]. In all data sets, each video sequence contains one action. The video sequences were converted into gray level to avoid the bias in color. The details of the data sets are summarized in Table 1 and some sample images from the video sequences are shown in Figure 2. In the experiments, we implemented a baseline approach using the "bag-of-words" representation for comparison.

We set the parameters for the experiments empirically. The bandwidth for a smooth temporal weight, σ, is set to 5. The vocabulary size K, which is the cluster number in the k-means clustering algorithm, is set to 250. In our experiments, we observe that the performance does not improve much when the level of the spatial pyramid $L > 1$. Therefore, we use the setting of $K = 250$ and $L = 1$ or $L = 2$, which leads to a 1250-dimension or 5250-dimension vector for the key frame representation.

4.2 Experimental results

With the same experimental setting on facial expression data set as in [4], we trained on one subject under one of the two lighting conditions and tested on: (1) the same subject under the same illumination, (2) the same subject under different illumination, (3) a different subject under the same illumination, and (4) a different subject under different illumination. The recognition rates in each scenario from Dollar's implementations[4], which is the baseline "bag-of-words" approach, and from our approaches are shown in Table 2. We can see that in the first scenario, the recognition task is easy. The baseline algorithm already achieved very high recognition rate, therefore our approaches only slightly improved the results. For the rest of the cases, our approaches with both configurations have demonstrated significant improvements.

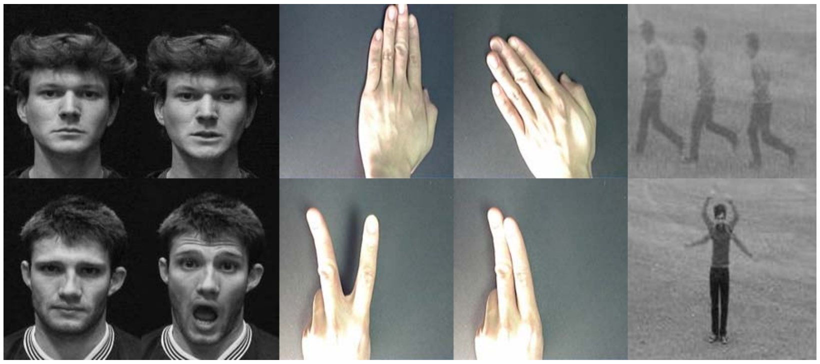


Figure 2: Sample images from the experiment data sets.

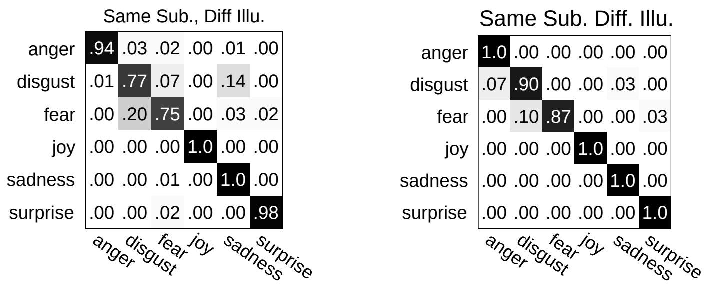
Methods	same sub. same illu.	same sub. diff. illu.	diff sub. same illu.	diff sub. diff illu.
Baseline	98.83	90.46	58.67	47.71
Our method (L=1)	**100.00**	**96.25**	74.38	71.71
Our method (L=2)	98.37	93.13	**74.83**	**73.25**

Table 2: The facial expression recognition rates(%) in different scenarios from the baseline "bag-of-words" algorithm and from our spatiotemporal representations with different spatial levels.

As an example, the confusion matrices of the six facial expressions in the second scenario from the baseline implementation and our spatiotemporal representation with L=1 are reported in Figure 3. It shows improvements on recognition rates in every category of the facial expressions.

From the experiment on facial expression data set, we do not observe any significant increase in performance beyond 1-level spatial pyramid configuration. This is because when $l = 1$, the 4 subdivisions of the key frame already roughly capture the sets of feature points' locations in the spatial domain while maintain tolerance for the locations variance in each cell. With more levels, the number of features points falling into each cell will be decreased, so the histograms might not be a good approximation for the feature's distribution. So we will use $l = 1$ for the rest of the experiments.

With leave-one-out cross-validation experimental setting, we tested the baseline and our proposed methods on all data sets. The confusion matrices from our method with L=1 are shown in Figure 4.The average recognition rates for all data sets, compared with other published results, are reported in Table 3. This demonstrates that our approach improves the "bag-of-words" baseline model and outperforms most of the other known methods while approaching the best known result.



Same Sub., Diff Illu.

	anger	disgust	fear	joy	sadness	surprise
anger	.94	.03	.02	.00	.01	.00
disgust	.01	.77	.07	.00	.14	.00
fear	.00	.20	.75	.00	.03	.02
joy	.00	.00	.00	1.0	.00	.00
sadness	.00	.00	.01	.00	1.0	.00
surprise	.00	.00	.02	.00	.00	.98

Same Sub. Diff. Illu.

	anger	disgust	fear	joy	sadness	surprise
anger	1.0	.00	.00	.00	.00	.00
disgust	.07	.90	.00	.00	.03	.00
fear	.00	.10	.87	.00	.00	.03
joy	.00	.00	.00	1.0	.00	.00
sadness	.00	.00	.00	.00	1.0	.00
surprise	.00	.00	.00	.00	.00	1.0

Figure 3: Comparison of the confusion matrices in the second scenario on the facial expression data set. The left confusion matrix is from Dollar's implementation[4] and the right confusion matrix is from our spatiotemporal representation with L=1.

Methods:	Baseline	Our method	Wong [22]	Niebles [13]	Wang [20]
Facial Expressions	91.33	**94.83**	83.33	none	none
Hand Gestures	85.81	**96.22**	91.47	none	none
KTH Actions	81.51	91.17	83.92	81.50	**92.43**

Table 3: The average recognition rates (%) for facial expression, hand gesture and KTH human action data sets obtained from different algorithms.

5 Discussion

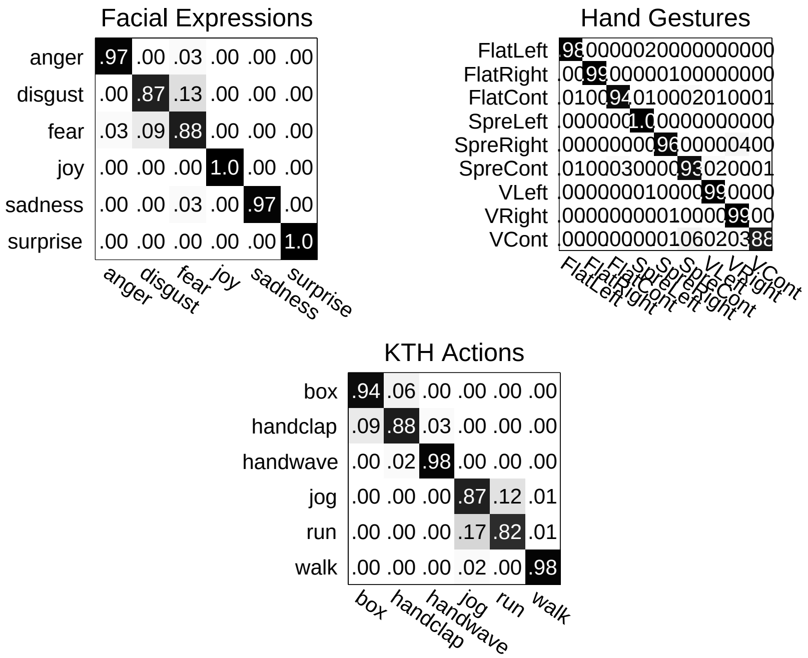
We have presented a spatiotemporal key frame representation for human action recognition. First, our approach selects the key frames of the video sequences based on their discriminative power. Next, our approach simultaneously integrates the spatiotemporal relations among local motion features with their appearance information and embeds these rich information in the representation for the selected key frames.

Our work differs from the pyramid match kernels[8, 10] in that: 1) Our goal is to find a suitable representation to integrate the spatiotemporal relations among motion features. The work in [8, 10] is seeking a suitable kernel function for two sets of image features. 2) Because our representation is a concatenated histogram, we measure the distance by Chi-square distance. The pyramid match kernels use histogram intersection as the distance function to satisfy the Mercer's condition.

In the future, we intend to further develop and disseminate this framework as a general method by automatically determining various hyper-parameters, which are currently empirically calculated.

Acknowledgments

This research is partially funded by NSF CAREER award IIS-0546372.



Facial Expressions

	anger	disgust	fear	joy	sadness	surprise
anger	.97	.00	.03	.00	.00	.00
disgust	.00	.87	.13	.00	.00	.00
fear	.03	.09	.88	.00	.00	.00
joy	.00	.00	.00	1.0	.00	.00
sadness	.00	.00	.03	.00	.97	.00
surprise	.00	.00	.00	.00	.00	1.0

Hand Gestures

	FlatLeft	FlatRight	FlatCont	SpreLeft	SpreRight	SpreCont	VLeft	VRight	VCont
FlatLeft	.98	.00	.00	.02	.00	.00	.00	.00	.00
FlatRight	.00	.99	.00	.00	.01	.00	.00	.00	.00
FlatCont	.01	.00	.94	.01	.00	.02	.01	.00	.01
SpreLeft	.00	.00	.00	1.0	.00	.00	.00	.00	.00
SpreRight	.00	.00	.00	.00	.96	.00	.00	.04	.00
SpreCont	.01	.00	.03	.00	.00	.93	.02	.00	.01
VLeft	.00	.00	.00	.01	.00	.00	.99	.00	.00
VRight	.00	.00	.00	.00	.01	.00	.00	.99	.00
VCont	.00	.00	.00	.00	.01	.06	.02	.03	.88

KTH Actions

	box	handclap	handwave	jog	run	walk
box	.94	.06	.00	.00	.00	.00
handclap	.09	.88	.03	.00	.00	.00
handwave	.00	.02	.98	.00	.00	.00
jog	.00	.00	.00	.87	.12	.01
run	.00	.00	.00	.17	.82	.01
walk	.00	.00	.00	.02	.00	.98

Figure 4: With leave-one-out cross-validation experimental setting, the confusion matrices on all three data sets from our spatiotemporal representation with L=1.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures: Multilevel local coding for visual recognition. In *ECCV06*, pages I: 30–43, 2006.

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Comput. Vis. Image Underst.*, 73(3):428–440, 1999.

[3] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[5] Gy. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 634, Washington, DC, USA, 2003. IEEE Computer Society.

[6] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *ICCV03*, page 726, 2003.

[7] D. M. Gavrila. The visual analysis of human movement: A survey. *Comput. Vis. Image Underst.*, 73(1):82–98, 1999.

[8] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV05*, pages 1458–1465, 2005.

[9] N. Ikizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *HUMO07*, pages 271–284, 2007.

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR06*, pages 2169–2178, 2006.

[11] Marcin Marszaek and Cordelia Schmid. Spatial weighting for bag-of-features. In *CVPR06*, pages 2118–2125, 2006.

[12] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126, 2006.

[13] J.C. Niebles, H. Wang, and F.F. Li. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC06*, page III:1249, 2006.

[14] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *CVPR06*, pages 2161–2168, 2006.

[15] R. Polana and R.C. Nelson. Detecting activities. In *DARPA93*, pages 569–574, 1993.

[16] J.C. Niebles S. Savarese, A. Del Pozo and L. Fei-Fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing. Copper Mountain*, 2008.

[17] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR06*, pages 2033–2040, 2006.

[18] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR04*, pages III: 32–36, 2004.

[19] C. Thurau. Behavior histograms for action recognition and human detection. In *HUMO07*, pages 299–312, 2007.

[20] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *HUMO07*, pages 240–254, 2007.

[21] Jutta Willamowski, Damian Arregui, Gabriella Csurka, Christopher R. Dance, and Lixin Fan. Categorization nine visual classes using local appearance descriptors. In *IWLAVS*, 2004.

[22] S.F. Wong, T.K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *CVPR07*, pages 1–6, 2007.

[23] Zhipeng Zhao, Akshay Vashist, Ahmed M. Elgammal, Ilya B. Muchnik, and Casimir A. Kulikowski. Combinatorial and statistical methods for part selection for object recognition. *Int. J. Comput. Math.*, 84(9):1285–1297, 2007.