Contextual Statistics of Space-Time Ordered Features for Human Action Recognition

Piotr Bilinski and Francois Bremond
INRIA Sophia Antipolis, STARS team
2004 Route des Lucioles, 06902 Sophia Antipolis, France
firstname.surname@inria.fr

Abstract

The bag-of-words approach with local spatio-temporal features have become a popular video representation for action recognition. Recent methods have typically focused on capturing global and local statistics of features. However, existing approaches ignore relations between the features, particularly space-time arrangement of features, and thus may not be discriminative enough. Therefore, we propose a novel figure-centric representation which captures both local density of features and statistics of space-time ordered features. Using two benchmark datasets for human action recognition, we demonstrate that our representation enhances the discriminative power of features and improves action recognition performance, achieving 96.16% *recognition rate on popular KTH action dataset and* 93.33% *on challenging ADL dataset.*

1. Introduction

Automatic recognition of human actions has gained tremendous interest in recent years. Video surveillance, video data indexing, video retrieving, human-computer interaction or sport event analysis are just few of many applications, in which action recognition plays the main role. In recent years, various methods have been proposed and much progress has been made. However, due to enormous variations in visual and motion appearance of both people and actions, camera view point, occlusions, noise and enormous amount of video data, action recognition still remains a challenging problem.

Over the last few years, many different action recognition techniques have been proposed. Existing techniques could be divided into four categories. The first group of techniques uses silhouette or body contour information to represent an action [5, 2, 1, 13, 20]. Such techniques usually require precise algorithms, which is often difficult to achieve, especially in real-world videos. The second category contains methods analysing motion trajectories [25, 28, 11, 32]. This group of techniques usually requires either feature point or object tracking. Reliable tracking is also difficult to achieve due to illumination variability, occlusions, noise, low discriminative appearance or drifting problems. The third category of techniques uses local spatio-temporal features [16, 27, 14, 8, 22] which have recently become a very popular video representation method for action recognition. Local spatio-temporal features have shown very good performance in recognition of various action classes. They are able to capture both visual and motion appearance. They are robust to viewpoint and scale changes, they are easy to implement and fast to process. Moreover, they do not require either object localization or tracking and in addition they are robust to background clutter. Over the last few years, many different local interest point detectors (like Harris3D [16], Cuboid [6], Hessian [35] or Dense sampling [33]) and many spatio-temporal descriptors (like HOG [17], HOG3D [14], HOF [17], Cuboid [6] or ESURF [35]) have been proposed. One of the most commonly used descriptors in the literature showing a high performance over the various datasets [33] are: Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) descriptors [17]. The former describes the local visual appearance and the latter characterizes the local motion appearance of the interest point. These descriptors are usually applied with the Harris3D corner detector [16],

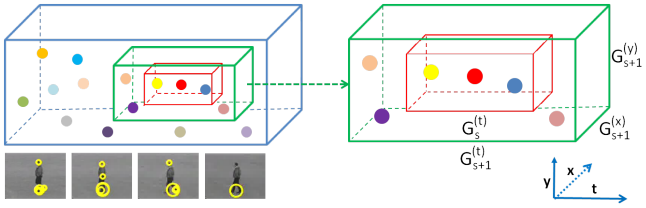


Figure 1: Multi-scale figure-centric neighbourhoods used to calculate Space-Time Ordered Contextual Features.

which is a space-time extension of the Harris operator.

Local spatio-temporal features are mostly used with a bag-of-words model. Together, they have shown to achieve high recognition rate across various datasets [21, 17, 26]. The bag-of-words model simplifies the structure of 3D video data assuming conditional independence across spatial and temporal domains. It encodes global statistics of features computing histogram of feature occurrences in a video. However, this technique has few limitations: it ignores relative position of features, local density of features, local pairwise relationships among the features and information about the space-time order of features. A common way to overcome these limitations is to use either spatio-temporal grids [17] or multi-scale pyramids [18]. Unfortunately, these methods are still limited in terms of detail description providing only a coarse representation. To overcome these limitations, techniques of the fourth category can be used to enhance the discriminative power of features and help in the task of action recognition. The fourth group of method is based on contextual information [7] extracting from the video content scene information [19, 23], spatio-temporal relations between trajectories [31], figure-centric features [34, 3, 15] or modelling human-object interactions [9]. Sun *et al*. [31] have proposed to model the spatio-temporal context of trajectories into transition matrix of a Markov process, and then extract its stationary distribution as the final context descriptor. The authors capture the local occurrence statistics of all types of trajectories within figure-centric neighbourhoods. Wang *et al*. [34] have proposed a representation that captures contextual interactions between interest points, based on the density of all features observed in each interest point's multiscale spatio-temporal contextual domain. However, both these methods are restricted to aggregate statistics over the video volume and ignore local pairwise relationships between features. Banerjee *et al*. [3] have proposed to learn the local neighbourhood relationships between local features, and train a CRF based human activity classifier. The neighbourhood relationships are modelled in terms of pairwise co-occurrence statistics. Kovashka *et al*. [15] have proposed to learn the shapes of space-time feature neighbourhoods that are the most discriminative for a given action category. The authors have proposed figure-centric statistics that capture the orientation between features. However, despite that these methods capture relationships between local features, they are still limited ignoring important information about the spatio-temporal order of features.

To differ from those ideas, we propose a novel representation based on quantized local spatio-temporal features. We enhance the discriminative power of features incorporating figure-centric information about the local spatio-temporal distribution and order of features. For each detected local feature, we define its neighbourhoods and compute statistics of pairwise co-occurring visual words within such neighbourhoods. Our representation captures not only, local density of features but also, local pairwise relationships among the features and information about the space-time order of features. We evaluate our approach on two publicly available datasets for human action recognition (KTH and ADL). We show that the proposed representation enhances the discriminative power of features and improves action recognition performance.

2. Proposed Approach

Local spatio-temporal features used with bag-of-words model have shown to achieve high action recognition performance. Recent methods have typically focused on capturing global and local statistics of features. However, existing approaches ignore relations between the features, particularly space-time arrangement of features, and thus may not be discriminative enough. Therefore, we propose a novel figure-centric representation which captures local density of features and local pairwise relationships among the features with information about the space-time order of features. The technique presented in this section enhances discriminative abilities of features and improves action recognition performance.

2.1. Interest Point Quantization

Firstly, we extract local interest points $\mathbb{P} = \{P_1, ..., P_n\}$ and their descriptors for each video sequence. Then, we cluster all the descriptors extracted from the training videos into k classes, called visual words. Finally, for each video sequence $\mathbb{V}$, we map the extracted interest points to the closest visual words using associated local descriptors:

$$\mathbb{V} = \{(P_1, c_1), \ldots, (P_n, c_n)\}, \quad (1)$$

where point P_i is represented as $P_i = [P_i^{(x)}, P_i^{(y)}, P_i^{(t)}]^T$ and c_i is an index of the closest visual word from the P_i.

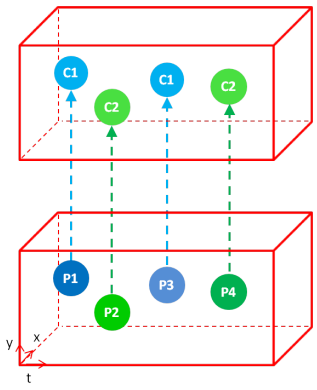
2.2. Figure-Centric Neighbourhoods

To specify Space-Time Ordered Contextual Features (STOCF), we initially define the neighbourhoods of detected points.

For each detected interest point P_i, we compute a set $\mathbb{S} = \{S_1, ..., S_{|S|}\}$ of multi-scale blocks around it. For simplicity, we define s-th scale neighbourhood of the point P_i as cuboid with side lengths $G_s^{(x)}$, $G_s^{(y)}$ and $G_s^{(t)}$ (Figure 1). The points that belong to such s-th scale cuboid can be defined as:

$$\mathbb{N}_{i,s} = \{P_j \in \mathbb{P} : \bigcap_{d \in \{x,y,t\}} |P_j^{(d)} - P_i^{(d)}| \leq W_s^{(d)}\}, \quad (2)$$

where $\forall_{d \in \{x,y,t\}} G_s^{(d)} = 2W_s^{(d)} + 1$.



$$\mathbb{N}_{3,s} = \{P_1, P_2, P_3, P_4\},$$

$$\mathbb{C} = \{C_1, C_2\},$$

$$\mathbb{N}_{3,s}^{(C_1)} = \{P_1, P_3\},$$

$$\mathbb{N}_{3,s}^{(C_2)} = \{P_2, P_4\},$$

$$\mathfrak{M}_{3,s}^{(F)} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{array} \begin{array}{c} \begin{array}{cccc} P_1 & P_2 & P_3 & P_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array},$$

$$\mathfrak{M}_{3,s}^{(C)} = \begin{array}{c} \\ C_1 \\ C_2 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & C_2 \end{array} \\ \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \end{array},$$

$$\mathfrak{CF}_{3,s} = [1, 3, 1, 1]^T,$$

Table 1: STOCF feature - sample of computation. The red cuboids represent a figure-centric s-th scale neighbourhood of the point P_3. The set of points belonging to this neighbourhood is marked as $\mathbb{N}_{3,s}$. $\mathbb{C}$ represents the set of visual words and $N_{3,s}^{(C_j)}$ is the set of points (belonging to the $\mathbb{N}_{3,s}$) that are assigned to the visual word C_j. An order of points is represented by binary matrix $\mathfrak{M}_{3,s}^{(F)}$, where $\mathfrak{M}_{3,s}^{(F)}(P_a, P_b) = 1$ means that point P_a occurs before point P_b. The matrix $\mathfrak{M}_{3,s}^{(C)}$ is obtained from corresponding points from the matrix $\mathfrak{M}_{3,s}^{(F)}$ using point to codebook mapping. Finally, the STOCF feature is marked as $\mathfrak{CF}_{3,s}$. Related elements of the matrices $\mathfrak{M}_{3,s}^{(F)}$ and $\mathfrak{M}_{3,s}^{(C)}$, and vector $\mathfrak{CF}_{3,s}$ are indicated by identical colours.

In the previous section, we have explained that each point is assigned to a certain visual word v. Therefore, we define $\mathbb{N}_{i,s}^{(v)}$ as a set of points in neighbourhood $\mathbb{N}_{i,s}$, which are assigned to the codebook element v:

$$\mathbb{N}_{i,s}^{(v)} = \{P_j \in \mathbb{N}_{i,s} : c_j = v\}, \quad (3)$$

where c_j is the index of the visual world assigned to the point P_j (defined in Equation 1).

2.3. Space-Time Ordered Contextual Features (STOCF)

Given computed local interest points and their neighbourhoods, we show how to compute the Space-Time Ordered Contextual Features (STOCF) for a specified point P_i and its neighbourhood $\mathbb{N}_{i,s}$.

We define STOCF as figure-centric statistics of features within 3D video patches. STOCF features are represented as histograms of pairwise co-occurring visual words. Each element of the histogram contains information about the relationship between two visual words - *e.g.* the value of x for pair of visual words (C_a, C_b) means that there is x pairs of points where the first point is assigned to the visual word C_a and the second to the C_b. More precisely, we compute non-negative matrix $\mathfrak{M}_{i,s}^{(C)}$:

$$\mathfrak{M}_{i,s}^{(C)} = \begin{bmatrix} \mathfrak{R}_{i,s}^{(1,1)} & \dots & \mathfrak{R}_{i,s}^{(1,k)} \\ \vdots & \ddots & \vdots \\ \mathfrak{R}_{i,s}^{(k,1)} & \dots & \mathfrak{R}_{i,s}^{(k,k)} \end{bmatrix}, \quad (4)$$

where $\mathfrak{R}_{i,s}^{(a,b)}$ is the cardinality of the set $\mathbb{R}_{i,s}^{(a,b)}$ ($\mathfrak{R}_{i,s}^{(a,b)} = |\mathbb{R}_{i,s}^{(a,b)}|$), and the set $\mathbb{R}_{i,s}^{(a,b)}$ represents pairs of co-occurring points which are organized in space-time:

$$\mathbb{R}_{i,s}^{(a,b)} = \{(P_j, P_k) \in \mathbb{N}_{i,s}^{(a)} \times \mathbb{N}_{i,s}^{(b)} : \left[\sum_{d \in \{x,y,t\}} w_d sgn(P_k^{(d)} - P_j^{(d)})\right] > 0\}, \quad (5)$$

Parameter w_d, which is explained later, is the weight for the dimension d and $sgn(x)$ is the signum function[1].

Finally, we define the STOCF features - $\mathfrak{CF}_{i,s}$, computed for the point P_i in the s-th scale neighbourhood $\mathbb{N}_{i,s}$, as matrix $\mathfrak{M}_{i,s}^{(C)}$ reshaped to a single dimensional vector:

$$\mathfrak{CF}_{i,s} = [\mathfrak{R}_{i,s}^{(1,1)}, \dots, \mathfrak{R}_{i,s}^{(1,k)}, \mathfrak{R}_{i,s}^{(2,1)}, \dots, \mathfrak{R}_{i,s}^{(k,k)}]^T, \quad (6)$$

If necessary, the size of the STOCF features can be reduced using *e.g.* PCA or LDA technique. However, due to the efficiency of small codebooks, methods for dimension reduction were not applied during our experiments. A calculation example of STOCF features is given in Table 1.

2.4. Action Recognition using STOCF features

To represent videos, we apply the bag-of-words model for each feature class (HOG-HOF and STOCF) independently. We construct visual vocabularies from training videos clustering computed features. Then, we assign each

[1]The signum function of a real number x is defined as follows: 1 if $x > 0$, 0 if $x = 0$, and -1 otherwise.

feature to its closest visual world. The concatenated histograms of visual world occurrences over video forms the final representation.

To classify a video, we use multi-class non-linear Support Vector Machines (SVM). We apply a χ^2 distance to compare two n-bins histograms $H_i = [H_i(1), ..., H_i(n)]^T$ and $H_j = [H_j(1), ..., H_j(n)]^T$:

$$\chi^2(H_i, H_j) = \frac{1}{2}\sum_{k=1}^{n}\left(\frac{(H_i(k) - H_j(k))^2}{H_i(k) + H_j(k)}\right) \quad (7)$$

This distance is then converted into SVM multi-channel χ^2 kernel using a multi-channel generalized Gaussian kernel:

$$K(H_i, H_j) = exp(-\frac{1}{A}\chi^2(H_i, H_j)) \quad (8)$$

where A is the normalization parameter set as in [17].

3. Experiments

Our experiments demonstrate the effectiveness of the proposed representation for a various of action categories. We evaluate our approach on two benchmark datasets for human action recognition - KTH and ADL datasets. Sample frames from video sequences of these datasets are presented in Figure 2. We show that using a small codebook (Section 2.1), we are able to enhance the discriminative power of features and improve action recognition performance.

3.1. Implementation Details

To detect interest points in a video, we use the sparse Harris3D corner detector [16]. We detect points in multiple spatial and temporal scales. Then, for each 3D video patch in the neighbourhood of a detected point, we compute HOG (Histogram of Oriented Gradient) and HOF (Histogram of Optical Flow) descriptors [17]. The detector and descriptors were selected based on their use in the literature and provide a good baseline for comparison with the state-of-the-art techniques. However, our action representation method is independent of the type of detector and descriptor, and can be used together with any other algorithm.

In order to quantize local features, we use the k-means clustering technique and nearest neighbour algorithm. To compute the bag-of-words representation, features are quantized to the codebook size of 1000, which has shown empirically to give good results. As a metric to calculate a distance between features and visual words, we use the L_2 norm. Since Harris3D corner detector calculates sparse spatio-temporal features, we set the weights w as $w^{(x)} = 1$, $w^{(y)} = 2$, $w^{(t)} = 4$ (Equation 5). To compute STOCF features, the HOG-HOF descriptors are quantized to small codebook sizes (10, 15, 20 and 25) and

Figure 2: Sample frames from video sequences of the KTH (first row) and ADL (second row) datasets.

figure-centric neighbourhoods are calculated for 8 different scales ($W_s^{(x)}$, $W_s^{(y)}$, $W_s^{(t)} \in \{4, 8, ..., 32\}$). The proper selection of neighbourhood size is important. Too small neighbourhood can contain only a few points and might not be discriminative. Too large volume may employ too many points and might also result in being not discriminative. Choosing an appropriate scale can be done in two ways: using Multiple Kernel Learning or cross-validation. In all our experiments, we calculate several codebooks to quantize local points (Section 2.1), and several multi-scale neighbourhoods to compute STOCF features. Then, we apply the cross-validation technique to both gauge the generalizability of the proposed approach, and select the most discriminative parameters. We use the Leave-One-Out Cross-Validation (LOOCV) technique, where videos of one person are used as the validation data, and the remaining videos as the training data. This is done repeatedly so that videos of each person are used once as the validation data.

3.2. KTH Dataset

The KTH [30][2] dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action is performed several times by 25 different subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). The dataset contains 599 video files. All sequences were recorded with 25 fps frame rate.

The dataset contains a set of challenges like: scale changes, illumination variations, shadows, different scenarios, cloth variations, inter and intra action class speed variations and low resolution (160×120 pixels spatial resolution).

We follow recent evaluations on the KTH dataset [21, 37, 12, 36, 10] using LOOCV scheme. In general, LOOCV assesses the performance of an approach with much more reliability than splitting-based evaluation schemes because it is much more comprehensive. Results from the exper-

[2] http://www.nada.kth.se/cvap/actions/

iments are presented in Table 2. Comparison of our approach with state-of-the-art methods in the literature using LOOCV technique is presented in the Table 3. For scenarios $s1$, $s2$, $s3$ and $s4$, our approach obtains the recognition rate of 98.67%, 95.33%, 92.62% and 98.00% respectively (selecting codebook size of 15, 25, 10 and 10 respectively). Overall, our approach obtains 96.16% recognition rate. The results clearly show that our representation enhances the discriminative power of features and outperforms state-of-the-art techniques.

Moreover, given quantized local interest points, we examine the average computation time of STOCF features using various codebooks (Section 2.1) and neighbourhoods. Results are presented in Table 4. We observe, that using small codebooks, STOCF features are very fast to calculate and achieve high action recognition rate.

s1	s2	s3	s4	**s1-s4**
98.67%	95.33%	92.62%	98.00%	**96.16%**

Table 2: KTH dataset: Evaluation of STOCF features. The table shows the action recognition rate.

Method	Year	Accuracy
Liu *et al*. [21]	2009	93.8%
Wu *et al*. [37]	2011	94.5%
Kim *et al*. [12]	2007	95.33%
Wu *et al*. [36]	2011	95.7%
Lin *et al*. [10]	2011	95.77%
Our method		**96.16%**

Table 3: KTH dataset: Comparison of our approach with state-of-the-art methods in the literature.

	Neighbourhood		
Codebook	4	12	20
10	$1.82ms$	$3.09ms$	$5.63ms$
20	$1.87ms$	$3.22ms$	$6.36ms$

Table 4: KTH dataset: Average computation time of STOCF features using various codebooks and neighbourhoods ($W^{(x)} = W^{(y)} = W^{(t)}$).

3.3. ADL Dataset

The ADL (University of Rochester Activities of Daily Living) [25][3] dataset contains ten types of human activities of daily living selected to be useful for an assisted cognition task. The full list of activities is: answering a phone, dialling a phone, looking up a phone number in a telephone directory, writing a phone number on a whiteboard, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware. Each action is performed three times by five different people. In total, the dataset contains 150 video sequences recorded with 30 fps frame rate and 1280×720 pixel resolution. The videos were down-sampled to the spatial resolution 640×360 pixels.

The dataset contains a set of challenges like: different shapes, sizes, genders and ethnicities of people, and difficulty to separate activities on the basis of a single source of information (e.g. eating banana and eating snack or answering a phone and dialling a phone).

The results from the experiments, together with a comparison of our approach with state-of-the-art methods in the literature, are presented in Table 5. Our approach obtains 93.33% recognition rate, which is comparable to [34] (where the authors use more efficient but more time-consuming and complex learning algorithms; also, [34] reaches 93.8% accuracy for the KTH dataset while our approach achieves 96.16% accuracy) and outperforms all other state-of-the-art methods [24, 29, 4, 28, 25]. The results clearly show that our representation enhances the discriminative power of features and improves the action recognition performance.

Method	Year	Recognition Rate (%)
Matikainen *et al*. [24]	2010	70%
Satkin *et al*. [29]	2010	80%
Banabbas *et al*. [4]	2010	81%
Raptis *et al*. [28]	2010	82.67%
Messing *et al*. [25]	2009	89%
Wang *et al*. [34]	2011	96% (93.8% for KTH)
Our method		**93.33%**

Table 5: ADL dataset: Comparison of our approach with state-of-the-art methods in the literature.

4. Conclusions and Future Work

We have proposed a novel figure-centric representation which captures statistics of space-time ordered features. This representation has been evaluated on two public benchmark datasets for human action recognition. We have obtained 96.16% recognition rate for the KTH dataset and 93.33% for the ADL dataset. This shows that our approach enhances the discriminative power of features and improves action recognition performance. In the near future work, we intend to examine more efficient learning algorithms

[3] http://www.cs.rochester.edu/~rmessing/uradl/

(like Multiple Kernel Learning) to combine STOCF features from different neighbourhoods, and to evaluate our representation using various interest point detectors and descriptors. We also intend to examine the proposed representation in a hierarchical manner, *i.e*. using STOCF features as descriptors in the interest point quantization process.

Acknowledgements.

This work was supported by the Région Provence-Alpes-Côte d'Azur. However, the views and opinions expressed herein do not necessarily reflect those of the financing institutions.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: a review. *CVIU*, 1999.

[2] M. Ahad, J. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 2010.

[3] P. Banerjee and R. Nevatia. Learning neighborhood cooccurrence statistics of sparse features for human activity recognition. In *AVSS*, 2011.

[4] Y. Benabbas, A. Lablack, N. Ihaddadene, and C. Djeraba. Action recognition using direction models of motion. In *ICPR*, 2010.

[5] J. Davis. Hierarchical motion history images for recognizing human motion. In *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, in conjunction with ICCV*, 2005.

[7] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *CVIU*, 2010.

[8] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, 2009.

[9] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.

[10] Z. Jiang, Z. Lin, and L. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *PAMI*, 2011.

[11] M.-B. Kaaniche and F. Bremond. Gesture recognition by learning local motion signatures. In *CVPR*, 2010.

[12] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *CVPR*, 2007.

[13] T.-S. Kim and Z. Uddin. *Silhouette-based Human Activity Recognition Using Independent Component Analysis, Linear Discriminant Analysis and Hidden Markov Model*. InTech, 2010.

[14] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[15] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[16] I. Laptev. On space-time interest points. *IJCV*, 2005.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[19] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[20] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, 2009.

[21] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[22] J. Liu and M. Shah. Learning human action via information maximization. In *CVPR*, 2008.

[23] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[24] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. In *ECCV*, 2010.

[25] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[26] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[27] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *CVPR*, 2009.

[28] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *ECCV*, 2010.

[29] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010.

[30] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[31] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, 2009.

[32] H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011.

[33] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[34] J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In *CVPR*, 2011.

[35] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.

[36] S. Wu, O. Oreifej, and M. Shah. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In *ICCV*, 2011.

[37] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.