PAIR-WISE EVENT DETECTION USING CUBIC FEATURES AND SEQUENCE DISCRIMINANT LEARNING

Xiaoyu Fang[1], *Yonghong Tian*[1], *Yaowei Wang*[2], *Chi Su*[1], *Teng Xu*[1], *Ziwei Xia*[2], *Wen Gao*[1*]

[1]School of EE&CS, Peking University, Beijing, China
[2]Department of Electronic Engineering, Beijing Institute of Technology, Beijing, China

ABSTRACT

Event detection in crowded surveillance videos is a challenging yet important problem. This paper focuses on pair-wise events that involve the interaction of two persons (e.g., people embrace, meet or split) in crowded videos. To detect such an event accurately, we should build an effective representation model that can characterize the sequential properties of two persons' interaction. Towards this end, we propose a novel pair-wise event detection approach using cubic features and sequence discriminant learning. A video sequence is first partitioned into several spatio-temporal cubes, and multiple features (e.g., statistics of trajectories, bag of spatio-temporal interest points) are extracted on these cubes and then fused to form a cubic feature descriptor under multiple kernel learning (MKL) framework. After that, the SVM with dynamic time alignment kernel is used to infer the existence of an event in the video sequence. Experimental results show that the proposed approach achieves the encouraging performance on TRECVid SED dataset.

Index Terms— Cubic feature, event detection, surveillance

1. INTRODUCTION

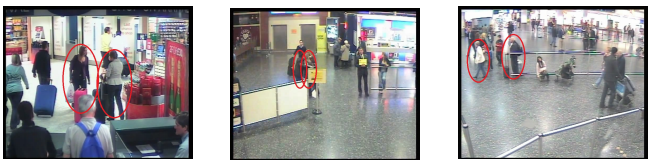
With the exponentially increasing deployments of surveillance cameras, one major challenge is how to automatically detect events of interest in surveillance videos. Unlike the laboratory environments, no assumptions have been made about the event instances occur in real-world surveillance scenarios, such as what kind of persons act the events, how the events are performed, and where the events happen. Various aspects of information about the observed persons should be utilized to discover special sequences of movements or interactions that indicate the happening of interest events.

This paper focuses on pair-wise event detection in a real surveillance dataset (i.e. TRECVid SED data corpus), which is collected from Gatwick Airport and provided by NIST [1]. Pair-wise events involve the interaction of at least two persons. Some samples of pair-wise events are shown in Fig.1. The detailed descriptions of these events given by NIST [1] are listed in Table 1. While analyzing pair-wise events, the persons' trajectories, motion and appearance are all elementary aspects. Meanwhile, sequential properties (e.g., the temporal order of motions in a period) are crucial to detect pair-wise events accurately. For example, two persons walk towards each other, and have a talk, that is PeopleMeet; two persons have a talk, and walk away from each other, that is PeopleSplitUp. BoW approaches form a histogram of features to represent events and discard the temporal order among these features. So the motion sequences "walking-and-talking" and "talking-and-walking" could not be distinguished using original BoW methods. To overcome the difficulties of the pair-wise event detection task, we propose a novel approach based on cubic feature and sequence discriminant learning method. First, each candidate video sequence is partitioned into a few spatio-temporal cubes. Then statistical trajectory descriptor and BoW interest point descriptor are extracted within each cube. And then these descriptors are linked across cubes and fused into one sequential cubic feature using MKL method. Moreover, a sequence discriminant learning method is employed to detect pair-wise events with sequential cubic features. The novelty and contributions of this paper are summarized as follows.

1. We design a novel cubic feature by partitioning a video sequence into several spatio-temporal cubes and exploiting multiple kernel learning (MKL) to fuse multiple features (i.e., statistical trajectory descriptors and BoW spatio-temporal interest points). The cubic features describe the sequential properties (e.g., temporal order of motions or interaction) of video events, and capture trajectory, motion and appearance characteristics of the objects performing the events.

2. We employ a sequence discriminant learning algorithm, namely SVM with dynamic time alignment kernel[2] (SVM^{dtak}), to detect pair-wise events with sequential cubic features. The SVM^{dtak} algorithm uses dynamic time alignment kernel to estimate the similarity of two sequences and takes temporal order of descriptors into

*This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No.61035001 and No.61072095, and National Basic Research Program of China under contract No.2009CB320906. Contact the author via yhtian@pku.edu.cn



(a) PeopleMeet (b) Embrace (c) PeopleSplitUp

Fig. 1. Samples of Pair-wise events.

Table 1. Description of three kinds of events: PeopleMeet, Embrace, PeopleSplitUp

Event	Description
PeopleMeet	One or more people walk up to one or more other people, stop, and some communication occurs.
Embrace	Someone puts one or both arms at least part way around another person.
PeopleSplitUp	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame.

consideration.

The rest of this paper is organized as follows. Section 2 reviews the related work of video event detection and human action recognition. In Section 3, the proposed pair-wise event detection approach is presented. Experimental results are reported and analyzed in Section 4. Finally, we conclude in Section 5.

2. RELATED WORK

Analyzing the relationships between the target object pairs is crucial while detecting pair-wise events. Zhou et al. [3] has confirmed that the trajectories give a global view of what is happening. A set of features based on the Granger Causality Test (GCT) is designed in [3] for describing the pair-activities. Hervieu et al. [4] proposes differential features combined with curvature and motion magnitude. Moreover, local appearance and movements indicate the interactions between persons. Spatio-temporal interest points [5, 6, 7] and local descriptors (i.e. HOG, HOF) have proved to be effective in describing local appearance and movements. Laptev in [5] extends the notion of spatial interest points into the spatio-temporal domain. They build on the idea of the Harris and Förstner interest point operators and detect local structures in space-time where the image values have significant local variations in both space and time. Dollar et al. [6] introduces a multidimensional linear filter detector, which detects denser interest points compared to Harris detector.

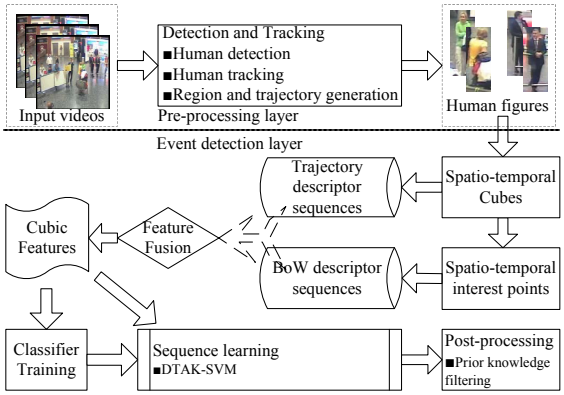


Fig. 2. Pair-wise video event detection using cubic feature and sequence discriminant learning.

Combining multiple features together [8, 9, 10] is regarded as an effective method to give more informative descriptions for videos. Sun et al. [8] extracts local descriptors and holistic features, and concatenates them together to represent actions. Bregonzio et al. [9] proposes a Clouds of Points (COP) feature, and combines it with BoW interest point feature using MKL to capture both distribution and appearance characteristics of the actions. Moreover, video events are considered as sequences of movements [11, 12] (e.g. walking and talking). The temporal order of these movements is discriminative in event detection. In [2], the dynamic time alignment kernel based support vector machine is proposed and has shown effectiveness in sequential-pattern recognition.

3. THE PROPOSED APPROACH

Our approach strives to capture the sequential properties of video events, and combines trajectory descriptors with BoW interest point descriptors. The input video sequence is pre-processed first using object detection and tracking algorithms. Therefore, candidate regions and trajectories are obtained. For a video subsequence in which an object pair coexists, we first partition it into k (variable) cubes, L (constant) frames as one cube. Then statistical trajectory descriptor is extracted and spatio-temporal interest points in candidate object's regions are detected within each cube. After that we cluster these interest points and generate a histogram descriptor for each cube according to a visual vocabulary built off-line with training points. So, the object pair is represented with a sequence of k trajectory descriptors and a sequence of k BoW. Then trajectory descriptor sequence and BoW descriptor sequence are fused into a cubic feature using MKL method. To classify pair-wise events with cubic features, the SVM with dynamic time alignment kernel (DTAK) is employed to handle sequential properties of features. Fig. 2 illustrates the flowchart of the proposed approach.

3.1. Cubic Feature

Let $VS = [I_1, ..., I_i, ..., I_T]$ be a candidate video sequence in which an object pair coexists. The regions of this object pair have been located and tracked using human detection and tracking algorithms. Here, we limit the max value of T to H frames (e.g., 1000). If a candidate video sequence exceeds H frames, the slide window of H frames is used to cut it into several shorter sequences. Then the candidate video sequence VS is partitioned into several spatio-temporal cubes. Each cube has fixed length of L frames (e.g., 10). Therefore, the i^{th} frame belongs to the $\lceil i/L \rceil^{th}$ cube, so the sequence could be represented as $VS = [I_1^1, ..., I_i^{\lceil i/L \rceil}, ..., I_T^{\lceil T/L \rceil}]$. Cubic feature is a sequence of spatio-temporal descriptors of these cubes. Descriptors of each cube include two sections: statistical trajectory descriptor and BoW interest point descriptor. The former describes statistic state of trajectories in current cube; The latter describes appearance and motion characteristics.

3.1.1. Statistical Trajectory Descriptor

Statistical trajectory descriptor describes relationships of the trajectory pair. Let $A_m = [a_1^1, ..., a_i^{\lceil i/L \rceil}, ..., a_T^{\lceil T/L \rceil}]$ and $B_n = [b_1^1, ..., b_i^{\lceil i/L \rceil}, ..., b_T^{\lceil T/L \rceil}]$ be motion trajectories of objects m and n, where a_i and b_i are tuples (x,y) of the object coordinates in 2D image plane at time i, and m, n are objects' identifiers. To represent the relationships of the objects in each cube, and remove the influences of occasional error caused by detection and tracking, statistical data is employed, such as mean distance one from another, mean relative speed magnitude, mean overlapped area of objects' regions. Meanwhile, the difference of these statistical data between current and next cubes is important as well. Therefore, trajectory descriptor of k^{th} cube is extracted as follows:

$$TD^k = \{c_{dis}^k, c_{sp}^k, c_{ov}^k, dc_{dis}^k, dc_{sp}^k, dc_{ov}^k\} \quad (1)$$

where c_{dis}^k, c_{sp}^k and c_{ov}^k are mean distance, mean relative speed magnitude and mean overlapped area within k^{th} cube respectively, and

$$\begin{cases} dc_{dis}^k = c_{dis}^{k+1} - c_{dis}^k \\ dc_{sp}^k = c_{sp}^{k+1} - c_{sp}^k \\ dc_{ov}^k = c_{ov}^{k+1} - c_{ov}^k \end{cases} \quad (2)$$

3.1.2. BoW Interest Point Descriptor

Ivan Laptev in [5] proposes a spatio-temporal interest point method which has proven to be effective for interpretation of visual events. A spatio-temporal image sequence can be modeled as a function $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$. The idea of spatio-temporal interest point detector is to find spatio-temporal locations where f has significant changes in both space and time domain. Laptev [5] extends the Harris corner function [13] defined for the spacial domain into spatio-temporal domain. Then spatio-temporal interest points of f can be found by detecting local positive maxima of the extended Harris corner function.

Around interest points, HOG (histograms of oriented gradients) and HOF (histograms of optical flow) descriptors are extracted to represent the appearance and motion characteristics. Then k-means algorithm is used to cluster HOG and HOF descriptors off line. And a BoW interest point descriptor for each cube is generated as a compact representation.

3.1.3. MKL Feature Fusion

The cubic feature includes two sequences of descriptors: a sequence of trajectory descriptors and a sequence of BoW interest point descriptors, and can be expressed as $X^* = [X^{Tr}, X^B]$, where X^{Tr} is the sequence of trajectory descriptors, and X^B is the sequence of BoW interest point descriptors. The trajectory descriptors and BoW interest point descriptors are generally independent and may not be equally informative in representing different events. Simply concatenating two descriptors within each cube could not get the optimal performance. An appropriate combination method of trajectory feature and BoW feature is necessary. We employ Multiple Kernel Learning (MKL) method for the feature fusion. MKL was first introduced by Bach et al. [14] to solve the problem of selecting the optimal combination of kernel functions for a specific feature for SVM. Recently, some researchers have adapt MKL to feature fusion task [15, 16, 9, 17], which could combine different features using a specific kernel function to achieve the optimal classification performance with SVM classifier.

In the original MKL method, the combined kernel function is expressed as: $K(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^{M} \beta_i K_i(\mathbf{x}, \mathbf{v})$, where $\beta_i \geq 0$ and $\sum_{i=1}^{M} \beta_i = 1$. To measure the similarity between a pair of cubic features, which include two sequences of descriptors, we employ a specific sequence kernel function K_s (Section 3.2), and estimate similarity of correspond sequences in cubic features respectively, and then combine them together. Therefore, the MKL based feature combination has the following form:

$$K_s^*(X^*, V^*) = \beta^{Tr} K_s(X^{Tr}, V^{Tr}) + \beta^B K_s(X^B, V^B) \quad (3)$$

The feature fusion task is to find a group of combination parameters (β^{Tr}, β^B) in Eq. 3 to optimize the classification performance of SVM using kernel K_s^*. This can be solved using generalized multiple kernel learning (GMKL) algorithm [18].

3.2. Sequence Discriminant Learning

Support Vector Machine (SVM) is one of the most successful statistical pattern classifier. However, basic SVM cannot

easily deal with the dynamic time sequence of features with different lengths. Several dynamic time warping (DTW) based kernel methods have adapted SVM to sequence processing [19, 2]. We employ the Dynamic Time Alignment Kernel (DTAK) [2] SVM to classify pair-wise events with cubic features.

Let $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$ be a sequence of vectors, where $\mathbf{x}_i \in R^n$, k is the length of the sequence, and the notation $|X|$ is used to represent the length of the sequence instead. Assume that we have two vector sequences X and V, and these two patterns may have different lengths. The dynamic time warping (DTW) algorithm is able to find the optimal path that minimizes the accumulated distance between two time series [20]. The DTW that is employed for SVM uses inner product or kernel function instead and finds the optimal path that maximizes the accumulated similarity.

$$K_s(X, V) = \max_{\psi,\theta} \frac{1}{W_{\psi\theta}} \sum_{i=1}^{N} w(i) K(\mathbf{x}_{\psi(i)}, \mathbf{v}_{\theta(i)}) \quad (4)$$

subject to

$$\begin{cases} 1 \le \psi(i) \le \psi(i+1) \le |X|, \psi(i+1) - \psi(i) \le Q \\ 1 \le \theta(i) \le \theta(i+1) \le |V|, \theta(i+1) - \theta(i) \le Q \end{cases} \quad (5)$$

where Q is a constant constraining the local continuity, ψ and θ stand for a warping path, N is the length of the warping path, $w(i)$ is a nonnegative weighting coefficient, $W_{\psi\theta} = \sum_i^N w(i)$ is a path normalizing factor, and K can be any conventional kernel function or simple inner product. In this paper, we set $K(\mathbf{x}_{\psi(i)}, \mathbf{v}_{\theta(i)}) = exp(-\gamma||\mathbf{x}_{\psi(i)} - \mathbf{v}_{\theta(i)}||^2)$, that is Radial Basis Function (RBF) kernel, and $w(i) = 1$, so $W_{\psi,\theta} = N$.

For the cubic feature, which includes a sequence of trajectory descriptors and a sequence of BoW descriptors, expressed as $X^* = [X^{Tr}, X^B]$, we consider the two sequences separately. Assume that we have two cubic features $X^* = [X^{Tr}, X^B]$ and $V^* = [V^{Tr}, V^B]$. we find the optimal paths of X^{Tr} with V^{Tr} and X^B with V^B respectively to maximize their similarity.

$$\begin{aligned} K_s^*(X^*, V^*) &= \beta^{Tr} K_s(X^{Tr}, V^{Tr}) + \beta^B K_s(X^B, V^B) \\ &= \beta^{Tr} \max_{\psi^{Tr},\theta^{Tr}} \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{x}^{Tr}_{\psi^{Tr}(i)}, \mathbf{v}^{Tr}_{\theta^{Tr}(i)}) \\ &+ \beta^B \max_{\psi^B,\theta^B} \frac{1}{M} \sum_{i=1}^{M} K(\mathbf{x}^{B}_{\psi^{B}(i)}, \mathbf{v}^{B}_{\theta^{B}(i)}) \end{aligned} \quad (6)$$

where K_s is the Dynamic Time Alignment Kernel (DTAK) represented as Eq. 4, K is RBF kernel, β^{Tr} and β^B are optimal combination parameters obtained in feature fusion step, (ψ^{Tr}, θ^{Tr}) and (ψ^B, θ^B) are warping paths of descriptor sequence pairs (X^{Tr}, V^{Tr}) and (X^B, V^B), N and M are lengths of the warping paths. Since the sequence kernel doesn't change the formulation of original SVM learning problem, the training algorithm for the original SVM can be used for sequence learning.

4. EXPERIMENTAL RESULTS

In our experiments, we choose a real surveillance dataset (i.e. TRECVid'08) [1] other than well controlled laboratory scenario videos (e.g., UT-Interaction [21]), because too many assumptions (e.g. clean backgrounds, no occlusions) exist in laboratory environments. The TRECVid'08 dataset is obtained from the Gatwick Airport which consists of 50-hour videos in the development set and 49-hour videos in the evaluation set. Ground truth annotations of the events occurring period are provided by NIST. The videos in development set are used for training and we further labeled the precise locations of persons performing the events. The videos in the evaluation set are used for testing. And persons in testing set are automatically detected and tracked using human detection and tracking algorithms. NIST formal toolkit[1] is used in our evaluation on experimental results. The parameters of the toolkit are kept the same with TRECVid'08 formal evaluation. Meanwhile, we participated TRECVid'12 SED task. Comparison results with some state of the art methods are reported in this section, too.

To quantitatively evaluate the performance, we use the Normalized Detection Cost Rate ($NDCR$) [1] as the primary measure. $NDCR$ is a weighted linear combination of the system's Missed Detection Probability (P_{Miss}) and False Alarm Rate (R_{FA}) (measured per unit time).

$$NDCR(S, E) = P_{Miss}(S, E) + Beta * R_{FA}(S, E) \quad (7)$$

where S is the evaluated system, E is the interest event and $Beta$ is composed of constant values that define the parameters of the surrogate application.

In the pre-processing layer, we first detect and track objects to obtain their locations and trajectories. The HOG [22] based human detector and Multiple Hypothesis Tracking (MHT) [23] method are applied in our system. The overall detection rate of human including both detection and tracking results is tuned to about 40% with the precision 79%.

4.1. Evaluation of Feature Fusion Performance

The combination parameters of trajectory descriptors and BoW interest point descriptors are first optimized using MKL method. The optimal parameter groups in our experiments are $(0.64, 0.36)$, $(0.78, 0.22)$, $(0.61, 0.39)$ respectively for PeopleMeet, Embrace, and PeopleSplitUp. Then, the fused cubic feature is compared with two single type of descriptors (Tr, BoW) and their concatenated form feature (Cat) on

[1] http://nist.gov/itl/iad/mig/tools.cfm

Table 2. Results of pair-wise event detection using different features. $Act.RFA$ and $Act.PMiss$ are the system's actual False Alarm Rate and Missed Detection Probability. $NDCR$ is Normalized Detection Cost Rate.

Meet	Act.RFA	Act.PMiss	NDCR
BoW	6.00	0.980	1.01
Tr	1.28	0.979	0.985
Cat	1.20	0.979	0.985
MKL	0.16	0.979	**0.980**
Embrace	Act.RFA	Act.PMiss	NDCR
BoW	1.62	0.988	0.996
Tr	0.22	0.975	0.976
Cat	4.22	0.948	0.969
MKL	5.40	0.908	**0.935**
SplitUp	Act.RFA	Act.PMiss	NDCR
BoW	0.32	0.994	0.996
Tr	24.60	0.863	0.986
Cat	24.80	0.863	0.987
MKL	30.96	0.797	**0.952**

TRECVid'08 testing set. The results are shown in table 2. It is proved that MKL method could find relatively appropriate parameters for feature fusion through optimizing classification performance. Note that the trajectory feature shows better results than BoW feature on testing set, and proves to be more robust. That is because scenarios in this dataset are full of occlusions. And there are a lot of noise among the interest points extracted within the objects' regions.

4.2. Evaluation of Sequence Discriminant Learning Performance

The sequence discriminant learning is unified framework including sequential form feature extraction and sequence kernel based training and classification. Because the cubic features extracted from different candidate videos may have different lengths, general SVM could not be directly used for classifying cubic features. To validate the effectiveness of sequence discriminant learning and compare it with general learning methods, we first set the number of cubes to 1, and apply RBF kernel SVM on this degraded cubic feature to obtain general learning results. Furthermore, we limit the number of cubes to 15 (an empirical value), and then use RBF kernel SVM to get another contrast result. The comparison results are listed in Table 3. Sequence discriminant learning method shows significant improvements comparing with both C1+SVM^{rbf} and C15+SVM^{rbf} cases. It is confirmed that setting the number of cubes to 1 discards the temporal orders of movements and get the worst performance. And cutting the sequences of descriptors into the same length will result in information loss, because event instances may span different lengths of time. Moreover, RBF kernel could not dynamic align the sequential features. So that movements represented by these features could not be appropriately aligned. Therefore, using SVM^{rbf}, video sequences could not be classified so well as using SVM^{dtak}.

Table 3. Comparison with fixed length feature and RBF kernel SVM using $NDCR$ measure.

Events	C1+SVM^{rbf}	C15+SVM^{rbf}	SVM^{dtak}
Meet	1.008	0.990	**0.980**
Embrace	1.008	0.994	**0.935**
SplitUp	1.017	0.983	**0.952**

Table 4. Comparison results with other methods on TRECVid'08 data corpus using $NDCR$ measure.

Event	Hauptmann [24]	Wilkins [25]	Ours
Meet	7.49	1.36	**0.980**
Embrace	2.74	1.27	**0.935**
SplitUp	4.85	-	**0.952**

4.3. Comparison Results with Other Methods

The proposed approach is compared with those best known methods on TRECVid'08 data corpus. In Table 4, the proposed approach shows a significant improvement over some state of the art methods using $NDCR$ measure. Meanwhile, we also participate TRECVid'12 SED task. Note that ground truth on TRECVid'12 dataset is not publically available. According to TRECVid'12 SED formal evaluation, comparison results are listed in Table 5. (Note that the results referenced in Table 4 and Table 5 are all best-performing ones reported in TRECVid'08 and TRECVid'12.) Our results are acceptable, with minimal $NDCR$ for PeopleMeet and comparable $NDCR$ with the best for Embrace and PeopleSplitUp. Note that our method is relatively stable comparing with the other two methods. The $NDCR$ of our system for all three kinds of events is lower than 1.

5. CONCLUSION

In this paper, we have proposed a novel approach based on cubic feature and sequence discriminant learning method to

Table 5. Comparison results with other methods in TRECVid 2012 SED tasks using $NDCR$ measure.

Event	CMU-IBM	MediaCCN	Ours
Meet	1.04	1.01	**0.980**
Embrace	**0.800**	0.955	0.951
SplitUp	**0.843**	0.984	0.978

detect pair-wise events in surveillance videos. The cubic feature is designed by partitioning a video sequence into a few spatio-temporal cubes and exploiting multiple kernel learning (MKL) to fuse statistical trajectory descriptors with BoW interest point descriptors. Cubic features describe the sequential properties of events and capture trajectory, motion and appearance information of the objects performing the events. In our experiments, the fused cubic features outperform statistical trajectory descriptors, BoW interest point descriptors and their concatenated feature. The sequence discriminant learning algorithm uses dynamic time alignment kernel to estimate the similarity of two sequences. The temporal order of movements is taken into consideration. Extensive experiments demonstrate that the proposed approach is more effective for pair-wise event detection than several existing methods and even achieves the encouraging performance comparable to the best results reported in the TRECVid'12 SED task.

6. REFERENCES

[1] NIST, "Trec video retrieval evaluation," `http://www-nlpir.nist.gov/projects/trecvid`.

[2] H. Shimodaira, "Dynamic time-alignment kernel in support vector machine," *Proc. Advances in Neural Information Processing Systems*, vol. 2, pp. 921–928, 2001.

[3] Y. Zhou, S. Yan, and T.S. Huang, "Pair-activity classification by bi-trajectories analysis," *CVPR*, 2008.

[4] A. Hervieu and P. Bouthemy, "Video event classification and detection using 2d trajectories," *VISAPP*, 2008.

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

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *VS-PETS*, 2005.

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

[8] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," *Computer Science Department*, 2009.

[9] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points," *Pattern Recognition*, 2011.

[10] Muhammad Muneeb Ullah, Sobhan Naderi Parizi, and Ivan Laptev, "Improving bag-of-features action recognition with non-local cues," *BMVC*, 2010.

[11] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," *ICCV*, 2007.

[12] Mohamed R. Amer and Sinisa Todorovic, "A chains model for localizing participants of group activities in videos," *ICCV*, 2011.

[13] Chris Harris and Mike Stephens, "A combined corner and edge detector," *Alvey Vision Conference*, pp. 147–152, 1988.

[14] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," *ICML*, 2004.

[15] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," *ICCV*, 2009.

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

[17] A. Noguchi and K. Yanai, "A surf-based spatio-temporal feature for feature-fusion-based action recognition," *ECCV*, 2010.

[18] Manik Varma and Bodla Rakesh Babu, "More generality in efficient multiple kernel learning," *ICML*, 2009.

[19] C. Bahlmann, "On-line handwriting recognition using support vector machines - a kernel approach," *IWFHR*, 2002.

[20] S. Salvador and P. Chan, "Fastdtw: Toward accurate dynamic time warping in linear time and space," *IOS Press*, pp. 561–580, 2007.

[21] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*, 2009.

[22] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," *CVPR*, 2005.

[23] D. Reid, "An algorithm for tracking multiple targets," *IEEE TAC*, vol. 24, pp. 843–854, 1979.

[24] A. Hauptmann, Robert V. Baron, Ming-Yu Chen, M. Christel, Wei-Hao Lin, L. Mummert, S. Schlosser, X. Sun, V. Valdes, and J. Yang, "Informedia @ trecvid2008: Exploring new frontiers," *In Online Proc. TRECVID 2008 workshop*, 2008.

[25] P. Wilkins, P. Kelly, C. Conaire, T. Foures, Alan F. Smeaton, and Noel E. O'Connor, "Dublin city university at trecvid 2008," *In Online Proc. TRECVID 2008 workshop*, 2008.