INRIA-WILLOW at TRECVid 2010: Surveillance Event Detection

Rachid Benmokhtar[1] and Ivan Laptev[2]

[1] IRISA/INRIA Rennes – Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France
rachid.benmokhtar@inria.fr
[2] INRIA Paris – Rocquencourt, 23 avenue d'Italie, 75013 Paris, France
ivan.laptev@inria.fr

Abstract. This paper presents a system evaluated in the Surveillance Event Detection (SED) task of TRECVid 2010 campaign. We investigate a generic statistical approach applied to seven event classes defined by the SED task. Our video representation is based on local space-time descriptors which are vector-quantized and aggregated into histograms within short temporal windows and spatial regions defined by the prior. We use priors on the spatial localization of actions estimated from the spatio-temporal annotation of actions in the training data. To recognize actions, we learn one-against-all action classifiers using non-linear SVMs. Each classifier is applied independently to localize temporal intervals of actions using window-scanning approach. We present results of six runs with variations in the two parameters: (i) classifier threshold and (ii) temporal extent of the scanning window.

1 Introduction

Automatic video surveillance holds a great potential for security applications. The large variability of video data with respect to view points, lighting, clothing of people as well as occlusions currently makes this task to a highly challenging research problem. To advance and to measure the progress in automatic video surveillance, TRECVID [1] provides Surveillance Event Detection (SED) task. The goal of SED is to evaluate detection of pre-defined event classes in real surveillance settings on the common video corpus and annotations provided to participants. In 2010 TRECVid provides a corpus with 144 hours of video from five video cameras located in the London Gatwick International Airport. The video corpus is annotated with the temporal extents and the labels for seven event classes: Pointing, PersonRunning, Embrace, ObjectPut, PeopleMeeting, PeopleSplitting, and CellToEar. In this paper we describe our system applied to the detection of all seven event classes. The paper is organized as follows: Section 2 presents our system architecture. Section 3 presents evaluation of the system on the validation set as well as the final automatic event detection results on the test set evaluated by NIST. Conclusions and future work plans are given in Section 4.

2 System Architecture

The overview of our event detection system is presented in figure 1. The system can be divided into 3 parts: (1) feature extraction and video representation, (2) learning event models, (3) temporal event localization.

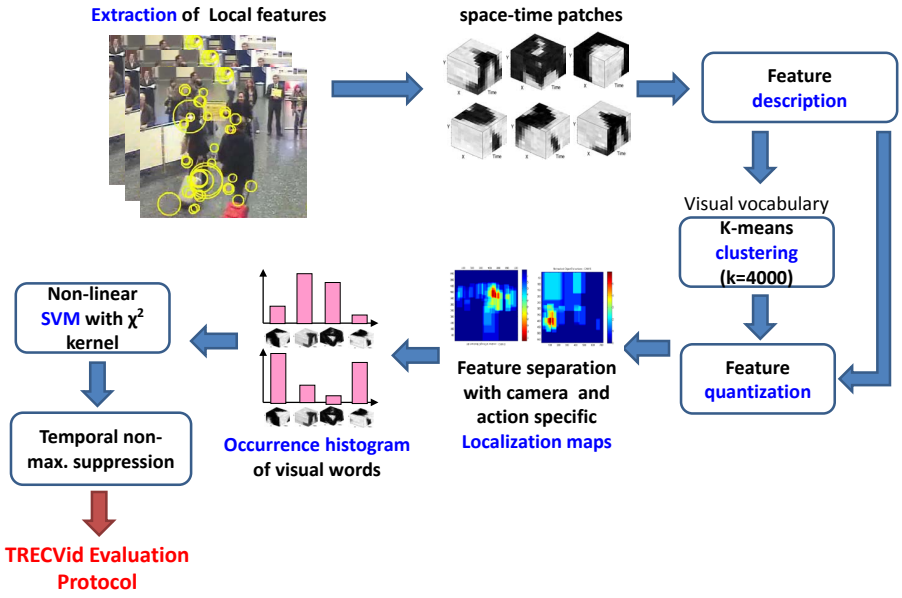

Fig. 1. Overall system for surveillance events detection.

2.1 Local space-time features

To represent events in the video, we use local spatio-temporal interest point detector (STIP) [2] combined with HOG and HOF descriptors [3]. STIP detector finds points in the video corresponding to locations with significant local variation of image values in both space and time. STIP detector in combination with HOGHOF descriptor has previously shown promising results for the task of action recognition and temporal action localization in realistic scenes [4, 3, 5].

To obtain a compact video representation, we vector-quantize STIP features. For this purpose we construct "visual vocabulary" [6] of local features by k-means clustering of a random subset of HOGHOF descriptors obtained from the training set. We set vocabulary size $K = 4000$ (number of visual words) which has empirically shown good results for a wide range of datasets. All features are assigned to their closest vocabulary word using Euclidean distance.

2.2 Localization maps

The main goal of this step is to use the camera-dependent prior position of each action in the scene. The SED TRECVid data are obtained from five static surveillance cameras in London Gatwick Airport. Each camera view represents a public scene (e.g. controlled access door, elevator close-up, etc). For training TRECVid provides only the temporal localization of actions without spatial information. We have noticed that the occurrence of specific actions is often biased towards specific locations in the scene. For example "Embrace" event is often located at the exit from the custom area, as illustrated in Figure 2.

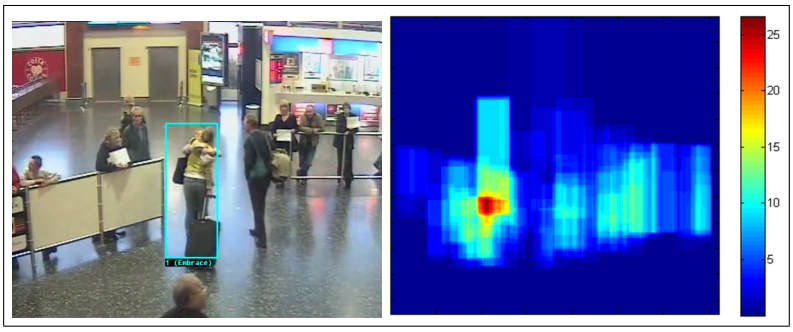

Fig. 2. (Left): Example of Embrace event for the camera 2; (Right): The localization map computed from all instances of Embrace action in the training data for this particular camera.

To construct spatial priors for events, we have annotated spatial extents of events in training videos by bounding boxes around people performing the action. For each video and event class we construct a map m by adding contributions $p = 1/n$ of all pixels within all annotated person bounding boxes in n frames. The camera and event-specific localization map $M^{c,e} = (\sum_i m_i^{c,e})/k$ is then obtained as an average of all k maps $m_i^{c,e}$, $i = 1..k$ for the specific camera c and event class e. Figure 3 illustrates localization maps obtained for events PersonRuns and ObjectPut and for the five cameras. As can be seen, the spatial localization of events varies both over the cameras and event classes. We use this information as a spatial prior for event recognition. For this purpose we obtain "action" and "no-action" regions by thresholding each map $M > \beta$ and separate local features w.r.t. action and non-action regions. We set threshold $\beta = 0.1$ empirically by cross-validation.

2.3 BoF representation

We represent each video interval by histograms of visual word obtained from the corresponding temporal window. For any camera c, histograms are computed and concatenated for action and non-action regions originating from seven action-specific localization maps $M^{c,e}$, $e = 1..7$ [7]. Each video is in this way represented by seven concatenated histograms regarded here as "channels". Segmenting local descriptors based on

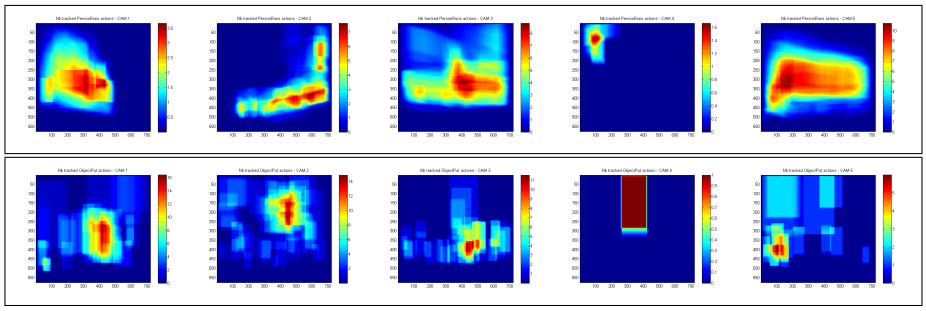

Fig. 3. The localization maps of PersonRuns and ObjectPut actions per camera view. The blue regions represent the low action activities areas and the red regions show the higher activities areas.

the foreground and background regions in video can be valuable in order to separate foreground features which are more likely to belong to the action from the background features which can help action recognition by capturing scene context.

An alternative separation of local features by spatio-temporal grids has been presented in [3]. Here we compare this approach with the one above. We test the spatial only grids defining six channels: 1x1 grid that corresponds to the standard BoF representation, 2x2, horizontal h3x1, vertical v1x3, denser 3x3 and center-focused o2x2 grids (see [3] for more details).

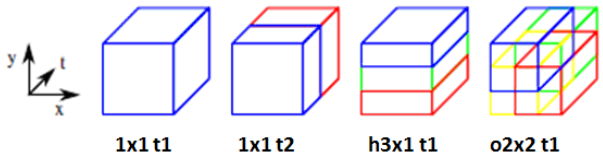

Fig. 4. Examples of spatio-temporal grids examples.

Experimental results in Table 1 demonstrate that the semantic decomposition of features according to action regions provides a significant performance improvement. LocMaps is compared to BoF and Grid approaches. We therefore use LocMaps results for the final submission.

2.4 Classification

To classify videos according to seven event classes, we use non-linear SVM classifier [8] with χ^2 RBF kernel applied on LocMaps histogram-based video representation. To combine evidence from seven event-specific channels, we use the product of kernels as in [9]. To train multiple classes of events, we use one-against-all approach.

Actions	BoF	Grid	LocMaps
PersonRuns	0.5247	0.6321	0.5714
ObjectPut	1.1108	1.1136	0.8889
CellToEar	0.4558	0.4198	0.1842
Embrace	0.2824	0.3151	0.1053
PeopleMeet	0.7165	0.6952	0.5035
PeopleSplitUp	0.5708	0.5694	0.3789
Pointing	0.271	0.271	0.1008
meanDCR	0.5617	0.5737	**0.3904**

Table 1. DCR scores per event using BoF, Grid and localization maps-based systems (the lower the score, the better the performance). The classification performance is obtained for the validation set.

To evaluate our system in terms of DCR scores, we fix the classifier threshold τ and declair an action to be present in videos with the SVM classification scores above τ. We use the values $\tau = \{-1, 0\}$ in the final submission (see Tables 2,3).

2.5 Events detection and temporal non-max. suppression

We use the temporal sliding window approach to accomplish the detection task. The sizes of the window used in our submission are $l_{win} = 100, 150, 200$ frames, we use sliding window step size of 10 frames. This creates a high number of windows in the test set ($\approx$ 390000 windows). To remove redundant detections, we apply temporal non-maximum suppression of detections based on their SVM classification scores.

3 Experimental results

Six runs are submitted based on two parameters variation: classifier threshold τ and scanning window length l_{win} :

- **Run 1 and 2:** we have fixed the parameters $l_{win} = 100$ and $\tau = \{-1, 0\}$ respectively,
- **Run 3 and 4:** $l_{win} = 150$ and $\tau = \{-1, 0\}$ respectively,
- **Run 5 and 6:** $l_{win} = 200$ and $\tau = \{-1, 0\}$ respectively,

The evaluation of our performance is based on the Detection Cost Rate (DCR) protocol using (F4DE) toolkit available from the MIG Tools Web page [3]. DCR is a single error measure that consists of a weighted linear combination of two errors: missed detections probability and false alarm rate. A lower DCR indicates better system.

In Table 2 we notice that Runs 1, 3, and 5 (computed with $\tau = -1$) have a high DCR comparing to the Runs 2, 4 and 6 respectively. This can be explained by the effect of $\tau = 0$ that results in the reduced false alarm rate. The false alarm rate is still fairly

[3] http://www.itl.nist.gov/iad/mig/tools/

high. A considerable portion of the false alarms are induced by the occlusion and the intersection among people. Some human actions appear similar to the true in terms of the motion patterns (e.g. touching hair is misclassified to CelltoEar, and it's very difficult to distinguish between ObjectPut and ObjectGet). Another common reason of false alarms is due to the low resolution of people that are far away from a camera.

In term of the window length, the effect depends on each action. For example PeopleSplitUp is a long time action that need bigger windows. In this case, Run 6 (i.e. $l_{win} = 300$) have an advantage comparing to the rest, contrary to Pointing that is a fast action where it needs a small window length. It's encouraging to note that some runs are competitive and able to get a min DCR close to 1 or lower with these fixed parameters. Table 3 presents results of our system compared to the best results obtained by all participants.

Actions	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
PersonRuns	3.795	**1.108**	3.814	1.127	3.824	1.118
ObjectPut	5.623	1.051	5.635	1.051	5.637	**1.043**
CellToEar	5.92	**1.113**	5.887	1.119	5.926	1.119
Embrace	11.112	**2.979**	11.13	2.991	11.112	2.985
PeopleMeet	13.325	3.927	13.365	**3.914**	13.386	3.921
PeopleSplitUp	12.847	3.168	12.847	3.157	12.876	**3.140**
Pointing	20.371	10.198	20.418	**10.197**	20.407	10.911

Table 2. Actual DCR scores by method and event.

Actions	Ranking	Best TRECVid sys.	Best INRIA sys.		
		DCR	DCR	FA	MissD
PersonRuns	5	0.737	1.108	472	102
ObjectPut	3	1.001	1.043	152	607
CellToEar	2	1.008	1.113	361	193
Embrace	6	0.967	2.979	7637	83
PeopleMeet	4	1.02	3.914	9532	345
PeopleSplitUp	5	0.959	3.14	7585	122
Pointing	6	0.999	10.197	28838	780

Table 3. Comparison between the best INRIA and TRECVid 2010 systems.

4 Conclusions and future work

In this paper we have described our first implementation and participation to SED TRECVid. A real surveillance dataset from London Gatwick airport have been ana-

lyzed, using spatio-temporal interest points descriptor and detector. The obtained performances show a good scores using this generic scheme, in particular for three actions: PersonRuns, ObjectPut and CellToEar. In the future work we plan to extend the current framework with better appearance models of actions as well as person-focused analysis of video.

Actions	Min.	Median	Mean	Max.
PersonRuns	9	54	64	240
ObjectPut	4	24	34	212
CellToEar	4	16	33	192
Embrace	5	71	144	3188
PeopleMeet	7	72	82	330
PeopleSplitUp	16	89	108	851
Pointing	3	22	38	1029

Table 4. Statistics of event duration in the number of frames.

5 Acknowledgements

We are grateful for financial support provided by the Quaero Programme, funded by OSEO.

References

1. A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.
2. I. Laptev, "On space-time interest points," *Int. J. of Comp. Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.
3. I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 2008.
4. O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video," in *Proc. Int. Conf. Comp. Vision*, 2009.
5. H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009.
6. J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," pp. 1470–1477.
7. M. Ullah, S. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues," in *British Machine Vision Conference*, 2010.
8. V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
9. J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. of Comp. Vision*, vol. 73, no. 2, pp. 213–238, June 2007.