Informedia@TRECVID 2011: Surveillance Event Detection

Longfei Zhang [1], Lu Jiang [2], Lei Bao [3], Shohei Takahashi [4], Yuanpeng Li [2]
Alexander Hauptmann [2]

[1] *School of Software, Beijing Institute of Technology, Beijing, 100081, P.R China*

[2] *School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA*

[3] *Laboratory for Advanced Computing Technology Research, ICT, CAS, Beijing 100190, China*

[4] *Graduate School of Global Information and Telecommunication Studies, WASEDA University, Tokyo, Japan*

Abstract: This paper presents a generic event detection system evaluated in the Surveillance Event Detection (SED) task of TRECVID 2011 campaign. We investigate a generic statistical approach with spatio-temporal features applied to seven event classes, which were defined by the SED task. This approach is based on local spatio-temporal descriptors, which is named as MoSIFT and generated by pair-wise video frames. Visual vocabularies are generated by cluster centers of MoSIFT features, which were sampled from the event part video clips. We also estimated the spatial distribution of actions by over generated person detection and background subtraction. Different slide window sizes and steps were adopted for different events by events' duration prior. Several sets of one-against-all action classifiers were trained using cascade non-linear SVMs and Random Forest, which could improve the classification performance in unbalanced data just like the SED datasets. 9 runs results were presented with variations in i) Slide window size ii) step size of BOW, iii) classifier threshold and iv) classifiers. The performance shows improvement over last year on the event detection task.

1. Introduction

Surveillance video recording is becoming ubiquitous in daily life for public areas such as supermarkets, banks, and airports. Thus it attracts more and more research interests and experiences rapid advances in recent years. A lot of schemes have been proposed for the human action recognition, among them, local interest points algorithm have been widely adopted. Methods based on feature descriptors around local interest points are now widely used in object recognition. This part-based approach assumes that a collection of distinctive parts can effectively describe the whole object. Compared to global appearance descriptions, a part-based approach has better tolerance to posture, illumination, occlusion, deformation and cluttered background. Recently, spatio-temporal local features [1-6] have been used for motion recognition in video. The key to the success of part-based methods is that the interest points are distinctive and descriptive. Therefore, interest point detection algorithms play an important role in a part-based approach.

The straightforward way to detect a spatio-temporal interest point is to extend a 2D interest point detection algorithm. Laptev et al. [2] extended 2D Harris corner detectors to a 3D Harris corner detector, which detects points with high intensity variations in both spatial and temporal dimensions. On other words, a 3D Harris detector finds spatial corners with velocity change, which can produce compact and distinctive interest points. However, since the assumption of change in all 3 dimensions is quite restrictive, very few point results and many motion types may not be well distinguished. Dollar et al. [7] discarded spatial constraints and focused only on the temporal domain. Since they relaxed the spatial constraints, their detector detects more interest points than a 3D Harris detector by applying Gabor filters on the temporal dimension to detect periodic frequency components. Although they state that regions with strong periodic responses normally contain distinguishing characteristics, it is not clear that periodic movements are sufficient to describe complex actions. Since recognizing human motion is more complicated than object recognition, motion recognition is likely to require with enhanced local features that provide both shape and motion information. Thus, MoSIFT feature [8] are proposed. MoSIFT detects spatially distinctive interest points with substantial motions by pair-wise frames. They first apply the well-know SIFT algorithm to find visually distinctive components in the spatial domain and detect spatio-temporal interest points with (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points.

However, in the local interest point algorithms, most of them [1-7] did not care where the interest points located, as their experiment scenes are relative simple and clear, and most of conditions, just one or two peoples have some actions**.** However, these conditions seldom hold in real-world surveillance videos. Even the same type of actions may exhibit enormous variations due to cluttered background, different viewpoints and many other factors in unconstrained real-world environment, such as TREC Video Retrieval Evaluation (TRECVID) [9]. To our knowledge, TRECVID has made the largest effort to bridge the research efforts and the challenges in real-world conditions by providing an extensive 144-hour surveillance video dataset recorded in London Gatwick Airport. In this dataset, the cameras are fixed, but the scenes are very complex, and there are a lot of people walking through on the scenes. Thus, if we just adopt the local interest points to detect the events on the scene, there are a lot of noise interest points for some events. In TRECVID 2011 Evaluation, there are 7 required events such as "CellToEar", "Embrace", "ObjectPut", "Pointing", "PeopleMeet", "PeopleSplitUp" and "PersonRuns". All of them are relative to the human. Therefore, we will use detection methods, such as human detection and foreground object subtraction, and tracking approaches to locate these interest points, and filter the noise interest points. Finally, we also adopt the results of human detection to estimate the correctness of detection.

In the following section, we will describe our system overview, and then the MoSIFT algorithm, the human detection and tracking will be introduced. After that, the experiments and discussions will be given. Finally, we will conclude the paper.

2. System Framework

For the tasks in TRECVID 2011 Event Detection Evaluation, we focus on human-related events. We mainly follow the framework we employed in TRECVID 2009 and 2010 Evaluation, which incorporates interesting point extraction, clustering and classification modules. In TRECVID 2009 Evaluation, the MoSIFT interesting points are extracted for each video firstly, and then Bag-of-Words (BoW) are adopted. After that, the cascade SVM will be trained. The details can be viewed in [16].

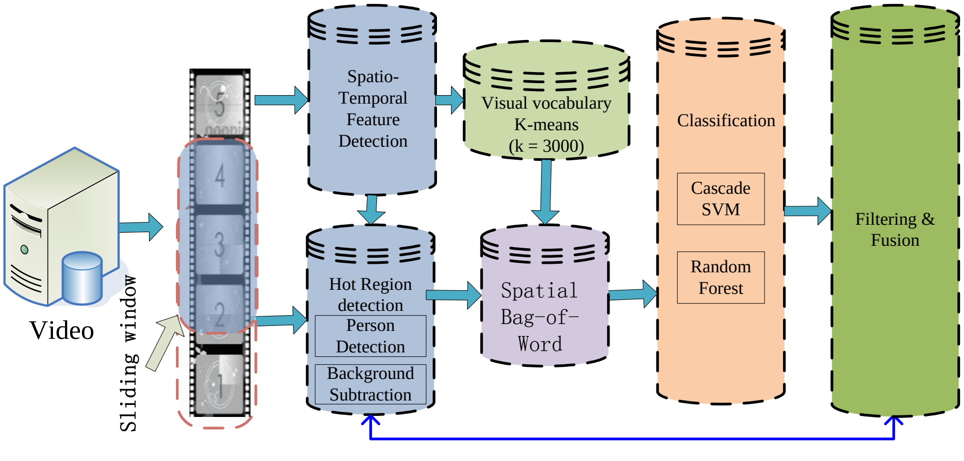


Figure 1: the framework of our surveillance event detection

In contrast, we extend our framework by three kinds of processing. Firstly, visual vocabularies were sampled from the event part video features. 3000 cluster centers were generated by K-means algorithm. In general, instance sampling from positive data might cause over fitting. But in surveillance video event detection, since the background did not change much, more positive samples might cause better video representation than random samples in visual vocabulary selection procedure. Secondly, each event had different durations. Laptev [18] tried several kinds of slide window size in Bag-of-Word procedure. We followed this idea and used three kind of slide window size for different event. Thirdly, we also used Spatial Bag-of-Word for a better video representation. Zhu [19] used a 3 layer spatial pyramid, 1x1 to 4x4, to represent the video. Laptev [18] used localization map to construct the BoW. Their experiment shows that localization map was better than grid spatial Bag-of-Word. However, localization map was based their additional annotation. Following the same idea, we used person detection and background subtraction analysis to find hot regions of actions automatically. Based on that region, we defined a different grid for video representation which will be introduced in subsection 3.4. Fourthly, for each frame, the MoSIFT points were extracted unsupervised. But these feature points might generate not by action but by noise, and we cannot discriminate them. Thus, the over generated human detection algorithm was adopted. We kept these MoSIFT points located in the region of human. Fifthly, both cascade SVM and Random Forest classifiers were trained for solving the unbalanced training and testing data. After got the probabilities of each event, we fused these results. The system framework is illustrated in the Figure.1.

3. MoSIFT Feature Based Action Recognition

For action recognition, there are three major steps: detecting interest points, constructing a feature descriptor, and building a classifier. Detecting interest points reduces the video from a volume of pixels to compact but descriptive interest points.

This section outlines our algorithm [8] to detect and describe spatio-temporal interest points. It was shown [8] to outperform the similar Laptev's method [2]. The approach first applies the SIFT algorithm to find visually distinctive components in the spatial domain and detects spatio-temporal interest points through (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points.

3.1. Motion Interest Point Detection and Feature Description

The algorithm takes a pair of video frames to find spatio-temporal interest points at multiple scales. Two major computations are applied: SIFT point detection [10] and optical flow computation matching the scale of the SIFT points.

SIFT was designed to detect distinctive interest points in still images. The candidate points are distinctive in appearance, but they are independent of the motions in the video. For example, a cluttered background produces interest points unrelated to human actions. Clearly, only interest points with sufficient motion provide the necessary information for action recognition.

Multiple-scale optical flows are calculated according to the SIFT scales. Then, as long as the amount of movement is suitable, the candidate interest point contains are retained as a motion interest point.

The advantage of using optical flow, rather than video cuboids or volumes, is that it explicitly captures the magnitude and direction of a motion, instead of implicitly modeling motion through appearance change over time.

Motion interest points are scale invariant in the spatial domain. However, we do not make them scale invariant in the temporal domain. Temporal scale invariance could be achieved by calculating optical flow on multiple scales in time.

After getting the MoSIFT interest points, we need describe these points. Appearance and motion information together are the essential components for an action classifier. Since an action is only represented by a set of spatio-temporal point descriptors, the descriptor features critically determine the information available for recognition.

The motion descriptor adapts the idea of grid aggregation in SIFT to describe motions. Optical flow detects the magnitude and direction of a movement. Since, optical flow has the same properties as appearance gradients, the same aggregation can be applied to optical flow in the neighborhood of interest points to increase robustness to occlusion and deformation.

The main difference to appearance description is in the dominant orientation. For human activity recognition, rotation invariance of appearance remains important due to varying view angles and deformations. Since our videos are captured by stationary cameras, the direction of movement is an important (non-invariant) vector to help recognize an action. Therefore, our method omits adjusting for orientation invariance in the motion descriptors.

Finally, the two aggregated histograms (appearance and optical flow) are combined into the descriptor, which now has 256 dimensions.

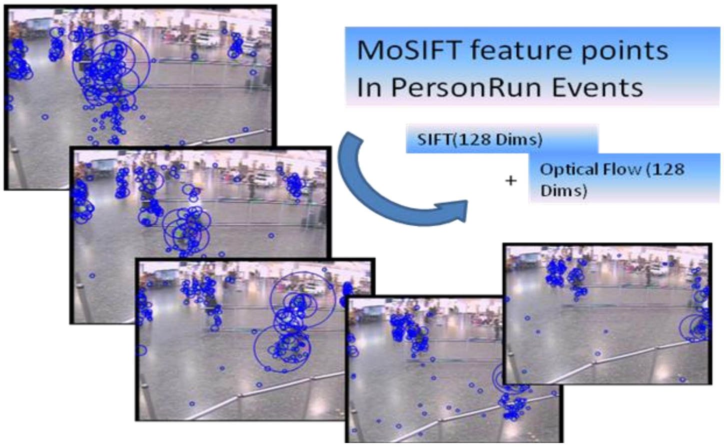


Figure 2: MoSIFT points in PersonRun Events. The center of blue circle is the location of the points. The radius of the blue circle is the scale of the movements

3.2. Hot Region Detection

MoSIFT feature does a great job in human behavior representation for human action recognition. However, Are the MoSIFT interesting points caused by human? The MoSIFT points might be caused by moving, light shaking, or shadow. If we could sample the MoSIFT points from human body or action region, we might reduce more noise interesting points. Thus, in this section, we use person detection and background subtraction method to create the hot region. These hot region could be used in feature selection and building spatial BoW.

Person detection is the most direct method to detect the area of human. Histogram of Oriented Gradient (HOG) feature [12] and Haar like feature [13] are the most popular features used in person detection. Locally normalized HOG descriptors are computed on a dense grid of uniformly spaced cells and use overlapping local contrast normalizations for improved performance. Haar like feature person detection used in VJ (Viola and Jones) works is using AdaBoost to train a chain of progressively more complex region rejection rules based on Haar-like wavelets and space-time differences. It consists of a filter that takes image windows from *n* consecutive frames as input, a threshold and a positive and negative vote. Since there are too many people in Gatwick surveillance video(especially camera 2, 3 and 5) , full body person detection is very limited in detecting the person blinded by some background objects, such as showed in figure 3. In our experiments, both HOG person upper/full body detectors and Haar person upper/full body detectors are trained on the development videos in Dev08 and INRIA dataset, and over generated for more reliable action region coverage.

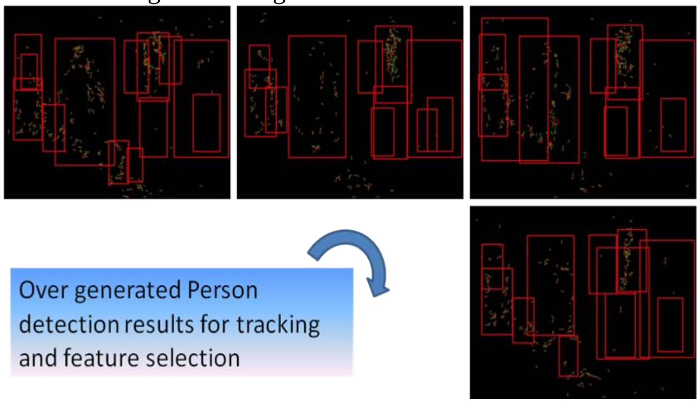

Over generated Person detection results for tracking and feature selection

Figure 3: Mosift points with over generated person detections results in sequence. The red rectangles are bounding boxes of objects detected by Person detecton, and the red points with green arrows are MoSIFT points. The green arrows are the direction of movements

We also use the background subtraction method to build the spatial priority maps. SED video data are captured from static cameras. People who don't act any movement have no relation with events. Therefore, background subtraction is effective to extract the area where people move and each event occurs.

To reduce the noise we use median filter, close operation and open operation for foreground. Foreground is expanded by open operation because surroundings area of foreground may be related with movements.

We built a spatial priority map by adding p= 1/n to each pixel with the foreground in *n* frames. Each spatial priority map is built for each camera. Fig4. shows spatial priority map from camera 2,3,4 and 5.

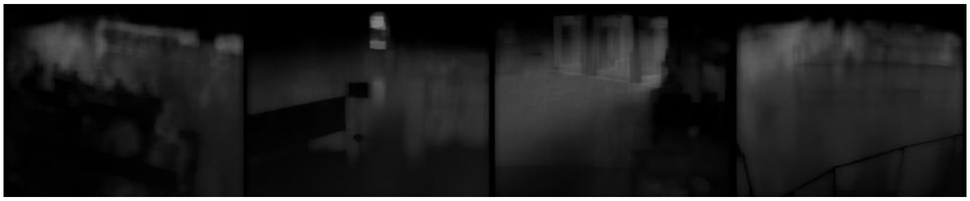


Figure 4: Priority map of cameras

3.3. Spatial Priority Maps based Spatial Bag of Features

From the prior knowledge of hot region, which were generated by person detection and spatial priority maps for each camera, the grid parsing strategy could be estimated. Based on this thinking, each frame is divided into a set of rectangular tiles or grids. The resulting Bow features are derived by concatenating the BoW features captured in each grid. Spatial bag of features allows for that the spatial distribution of interest points could be taken into account by the classifier. Generally, it facilitate the classifier distinguish events happen in different part of the frame. In our experiment, 8 rectangular tiles are adopted, as illustrated Fig.5. We choose to partition the frame in this way because according to our empirical observation in training set, the partition may reasonably capture the hot region of actions.

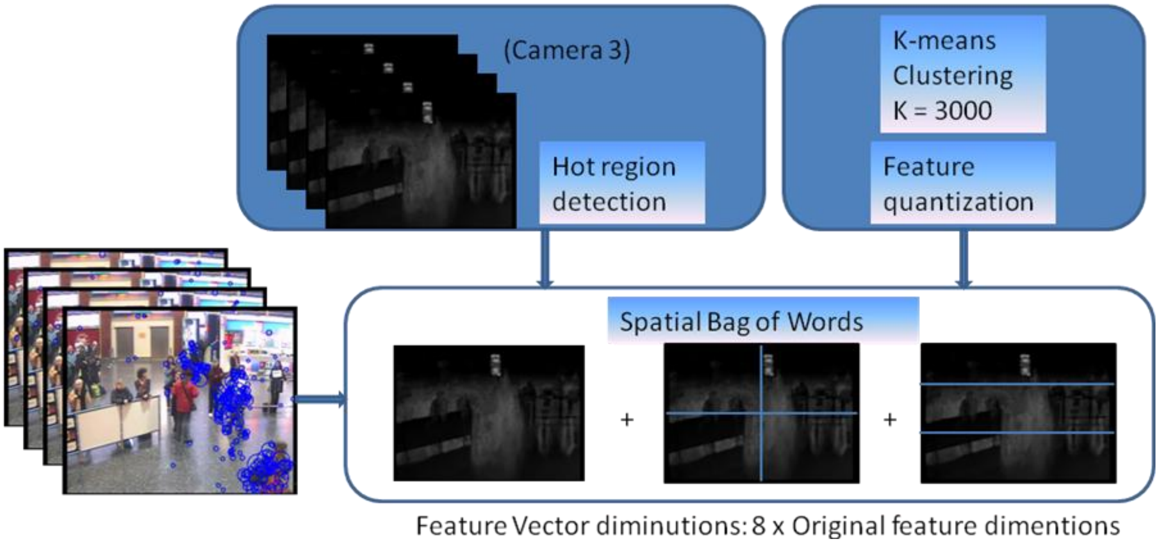

(Camera 3)
Hot region detection
K-means Clustering K = 3000
Feature quantization
Spatial Bag of Words
+
+
Feature Vector diminutions: 8 x Original feature dimentions

Figure 5: Spatial Bag-of-Words.

4. Experiments and Discussion

In TRECVID 2011 Event Detection Evaluation [9], 99 hours videos are provided as the development set and about 44 hours videos as the evaluation set, where the videos were captured using 5 different cameras with image resolution 720×576 at 25 fps.

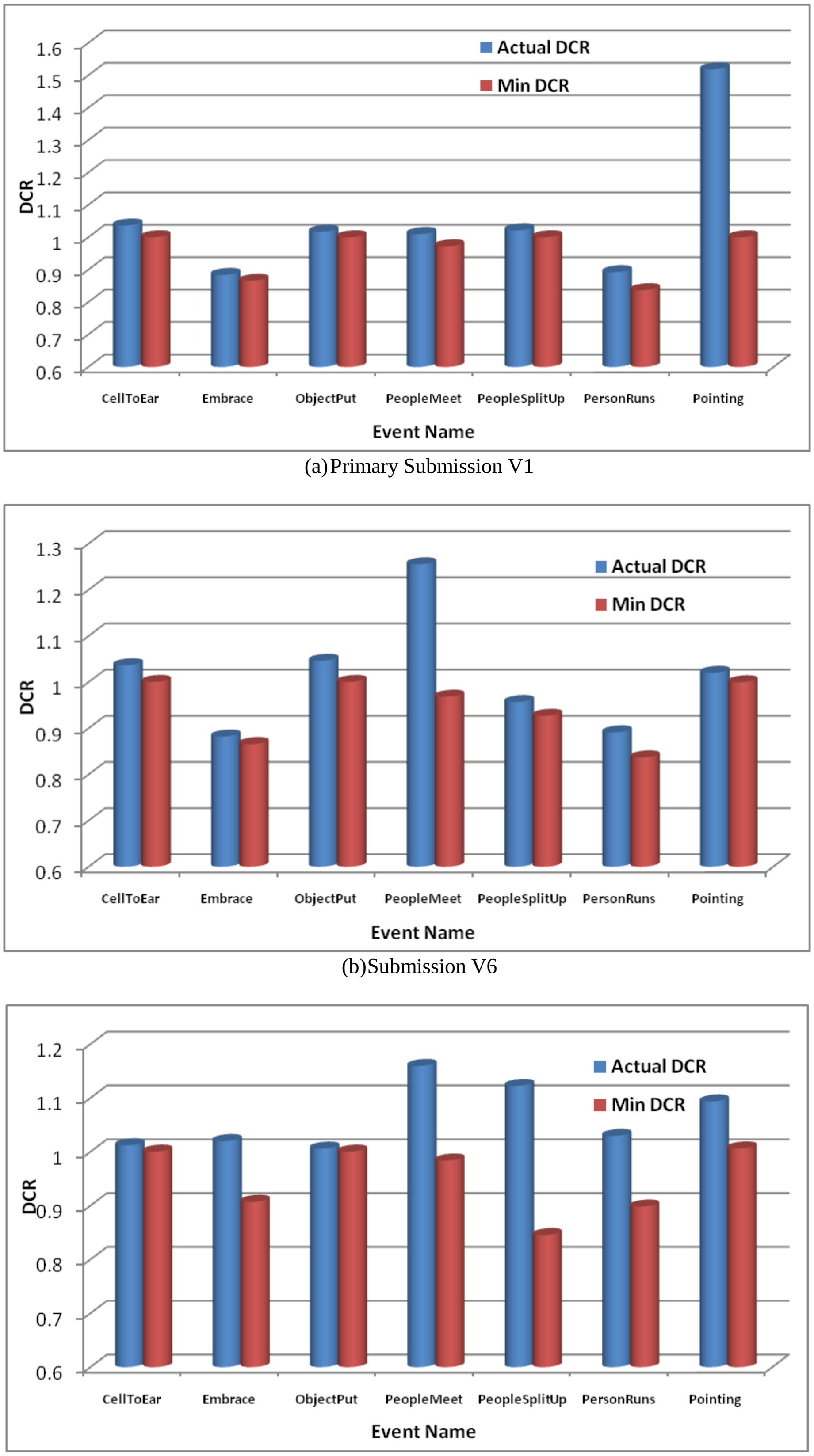
Our experiment start by extracting MoSIFT interest points in each sliding window. MoSIFT is a scale invariant local feature which is less affected by global appearance, posture, illumination and occlusion.

Figure 2 illustrates an example of the extracted MoSIFT features from Person Run Events. It can be seen that MoSIFT feature reasonably captures the areas with human activity. As different sliding window size can affect the final performance, in the experiment we manually set different window sizes for different events to ensure the window can capture the whole event. E.g. in our primary submission, for the event “ObjectPut”, “PeopleMeet” and “PeopleSplitUp”, the window size is set to 60 frames and repeats every 10 frames as these event cover a relatively long time-span; whereas for the event “PersonRuns”, “CelltoEar”, Embrace and Pointing , the window size is set to 60 and repeats every 15 frames.

We assume that an event can be described by a combination of different types of small motions. Therefore we use BoW to quantify MoSIFT feature to a fixed number vector feature of each key frame. We use K-means clustering to find the conceptual meaningful clusters and each cluster is treated as a visual word in BoW approach. All the visual words consist of a visual word vocabulary. Then interest points in each key frame are assigned to clusters in the visual vocabulary which are their nearest neighbors. In the end, each key frame is presented by a visual BoW features. In our experiments, the vocabulary size is $3,000$, and a soft boundary to form our bag-of-word features is applied. The spatial bag-of-word is also incorporated while constructing the resulting vocabulary. Each frame is divided into 8 tiles i.e. 1×1, 2×2 and 1×3 rectangular tiles as illustrated in Figure 5. Consequently the dimension of resulting BoW features is $8 \times 3,000 = 24,000$. Once the BoW features are obtained, a binary SVM [11] classifier with a χ^2 kernel is trained for each event. Finally we apply one-against-all strategy to construct action models.

The sliding window results in a highly unbalanced dataset (positive windows are much less frequent than negative windows). Two approached is conducted to tackle the unbalanced data. The first is cascade SVM. We build a one, five and ten layers cascade classifier to overcome this imbalance in the data and reduce false alarms. For each layer, we choose an equal ratio of (positive vs. negative) training data to build a classifier to favors to positive examples. This leads the classifier with high detection rates. In the training process, the cross-validation is adopted. By cascading five or ten layers of these high detection rate classifiers, we can efficiently eliminate a good amount of false positives without losing too many detections. We also aggregate consecutive positive predictions to achieve multi-resolution. The second approach is under sampling the majority class. Each sub-dataset is constructed in a way that all positive instances are preserved and negative instances are randomly supplemented. We choose negative to positive ratio 2.5 to reflect the natural imbalance in data. For each sub-dataset, we train a model using the random forest algorithm, in which the number of trees in forest is set to 100 and the maximum number of features is set to 2000. Finally all models are trained and aggregated together by averaging their votes. The empirical analysis suggests both solutions can improve the performance on unbalanced dataset.

Figure 6 summarizes the DCR(Detection Cost Rate) analysis for some of our this year’s final submissions. Fig.6 (a) illustrates our primary submission, in which the window size is set to 60 frames and repeats every 10 frames for the event “ObjectPut”, “PeopleMeet” and “PeopleSplitUp”. The window size is set to 60 and repeats every 15 frames for other events. Fig.6 (b) illustrates our submission version 6, in which the window size is set to 60 frames and repeats every 10 frames for all events. Fig.6 (c) illustrates our run 7th, which share the same sliding window setting with our primary submission. The difference lies in this version is that it adopts the random forest rather than the cascade SVM used in other submissions. It can be seen that in our primary submission (SVM) the Actual DCR (ADCR) and Minimum DCR (MinDCR) are quite similar, which is mainly because we search the best threshold in the training set. Generally, both cascade SVM classifiers and random forest classifiers are robust. Although, generally, cascade SVM classifiers outperform random forest classifiers, random forest classifiers are much faster and give relatively good predictions.



(a) Primary Submission V1

(b)Submission V6

(c) Submission V7

Figure 6: the DCR(Detection Cost Rate) Analysis results

Table 1 summarizes the comparison between this year's result and last year's result in terms of MinDCR, in which the first and the third row represents the best score of our 2010 and 2011 submission for each event, respectively; the second row indicates the best score for each event reported in TRECVID 2010 document[17]. It can be seen that, compared with our last year's result, we improve the performance for the event Embrace, "PeopleSplitUp", "PersonRuns" and "Pointing", in which "Embrace", "PeopleSplitUp" and "PersonRuns" even beats the last year's best results. The most significant improvement we achieve this year regards to the event "PeopleSplitUp", in which the MinDCR is reduced by 20.7%. The improvement is probably credited to the larger vocabulary and the introduction of spatial BoW. However, this strategy results in a considerable high dimension space, e.g. this year's 24,000 dimension versus last year's 2,000. Therefore efficient algorithms are recommended to be applied in this considerable high dimension space, which explains the reason of adopting cascade SVM and random forest algorithm in our experiment.

Table 1: Comparison between the best result of this year and last year in MinDCR

	CellToEar	**Embrace**	**ObjectPut**	**PeopleMeet**	**PeopleSplitUp**	**PersonRuns**	**Pointing**
2010 CMU	1.0003	0.9838	1.0003	0.9793	0.9889	0.9477	1.0003
2010 Best	**1**†	0.9663†	**0.9971**●	0.9787‡	0.9889	**0.6818**⋆	**0.996**†
2011 CMU	1.0003	**0.8658**	1.0003	**0.9684**	**0.7838**	0.837	0.9996

†the result attributes to IPG-BJTU ‡ the result attributes to TJU ⋆ the result attributes to QMUL-ACTIVA ● the result attributes to CRIM. The best score for each event is in bold

Table 2 presents the comparison between CMU results with TRECVID 2011 best results. The comparison is conducted on each team's primary submission. The "Best TRECVid Sys. MinDCR" column denotes for the best MinDCR reported in TRECVID 2011 Formal Evaluation Comparative Results. The ranking column represents our results' ranking among all groups in terms of the MinDCR.

Table 2: Primary run results comparison between CMU and TRECVID 2011 best results

Actions	**Ranking**	**Best TRECVid Sys.**	**CMU sys. Primary Run**				
		MinDCR	**MinDCR**	**ADCR**	**#CorDet**	**#FA**	**#Miss**
CellToEar	1	**1.0003**	**1.0003**	1.0365	1	127	193
Embrace	1	**0.8658**	**0.8658**	0.8840	58	657	117
ObjectPut	4	0.9983	1.0003	1.0171	0	57	620
PeopleMeet	1	**0.9724**	**0.9724**	1.0100	45	336	404
PeopleSplitUp	5	0.8809	1.0003	1.0217	3	115	184
PersonRuns	1	**0.8370**	**0.8370**	0.8924	26	413	81
Pointing	3	0.9730	1.0001	1.5186	132	1960	931

5. Conclusion

In this paper we have described our implementation to SED TRECVID2011. A real surveillance dataset from London Gatwick airport have been analyzed, using spatio-temporal interest points descriptor, MoSIFT, and spatial feature selection and representation. The obtained performances show good scores using this generic scheme, in particular for three actions: "PersonRuns", "PeopleMeet" and "Embrace". In the future work we plan to extend the current framework with better spatio-temporal models of actions as well as person-focused analysis of video.

6. Acknowledgments

This work was supported by the Nation Science Foundation under Grant No. IIS-0205219 and Grant No. IIS-0705491.

Reference

[1]. Schuldt, C. Laptev, and B. I. Caputo. Recognizing human actions: a local SVM approach. ICPR(17), pp 32-36, 2004.

[2]. I. Laptev and T. Lindeberg. Space-time interest points. ICCV, pages 432–439, 2003.

[3]. S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. ICCV, pp 1-8, 2007.

[4]. A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. BMVC, 2008.

[5]. G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. ECCV, pp 650-663, 2008.

[6]. A. Oikonomopoulos, L. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. ICME, pp 1-4, 2005.

[7]. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp 65-72, 2005.

[8]. M.-y. Chen and A.Hauptmann. MoSIFT: Reocgnizing Human Actions in Surveillance Videos . CMU-CS-09-161, Carnegie Mellon University, 2009.

[9] National Institute of Standards and Technology (NIST): TRECVID 2009 Evaluation for Surveillance Event trecvidDetection.http://www.nist.gov/speech/tests/trecvid/2009/ and trecvidhttp://www.itl.nist.gov/iad/mig/tests/trecvid/2009/doc/eventdet09-evalplan-v03.htm, 2009. 1, 7

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[11] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[12] N. Dalal and B. Triggs. Histogram of oriented gradient for human detection. In CVPR, 2005.

[13] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In CVPR, 2003

[14] Medeiros, H., Park, J.; Kak, A.A parallel color-based particle filter for object tracking, In CVPR2008 Workshops, p1-8, June 2008

[15] Babenko, B. Ming-Hsuan Yang; Belongie, S. Visual tracking with online Multiple Instance Learning. In CVPR2009 Workshops), p 983-90, 2009.

[16] Ming-yu Chen, Huan Li, and Alexander Hauptmann. Informedia @ TRECVID 2009: Analyzing Video Motions.

[17] Over, Paul and Awad, George and Fiscus, Jon and Antonishek, Brian and Michel, Martial and Smeaton, Alan F. and Kraaij, Wessel and Qu énot, Georges (2011) TRECVID 2010 – An overview of the goals, tasks, data, evaluation mechanisms, and metrics. In: TRECVID 2010, 15-17 November 2010, Gaithersburg, Md., USA.

[18] R. Benmokhtar and I. Laptev, “INRIA-WILLOW at TRECVid2010: Surveillance Event Detection”, http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/inria-willow.pdf

[19] G. Zhu, M. Yang, K. Yu, W. Xu and Y. Gong, “Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor” In proc. ACM Multimedia 2009, Beijing.