BIT @ TRECVid 2013: Surveillance Event Detection

Yicheng Zhao[1], Binjun Gan, Shuo Tang, Jing Liu, Xiaoyu Li,
Yulong Li, Qianqian Qu, Xuemeng Yang, Longfei Zhang[2]

Key Laboratory of Digital Performance and Simulation Technology
School of Software, Beijing Institute of Technology, Beijing, 100081, P.R China

{ [1]1120102071 , [2]longfeizhang }@bit.edu.cn

Abstract: In this paper, we present an event detection system evaluated in TRECVid 2013. We investigated a generic statistical approach with spatio-temporal features applied to the event "Object Put". This approach is based on local spatial constrained spatio-temporal descriptors, named as SC-MoSIFT. We extended the spatio-temporal features, MoSIFT, by relative position to camera. We also statistic the frequency and the location of each event occurs, and use selective hot region to construct spatial Bag-of-Feature. Non-linear SVM is exploited to train classifier for each event on each camera. In the limited experimental time, we adopted experiments and got comparable results to show the effectiveness of our approach.

1. Introduction

This year we submitted 2 interactive Event Detection Evaluation runs on event "Object Put".

Table 1 Runs

Run ID	Description
BIT _1	Origin annotation, MoSIFT [1], visual vocabulary size = 3000, spatial BoF, cascade SVM with Chi-Square kernel
BIT_2	Enhenced annotation, SC-MoSIFT, visual vocabulary size = 3000, action based spatial BoF, SVM with Chi-Square kernel [2]

2. System Framework

For the tasks in TRECVid 2013 [3] Event Detection Evaluation, we focus on one event "Object Put". Our primary run (BIT_1) use the framework CMU employed in TRECVID 2011 [4], which incorporates interesting point extraction, clustering and classification modules. We use BIT_1's result as our baseline.

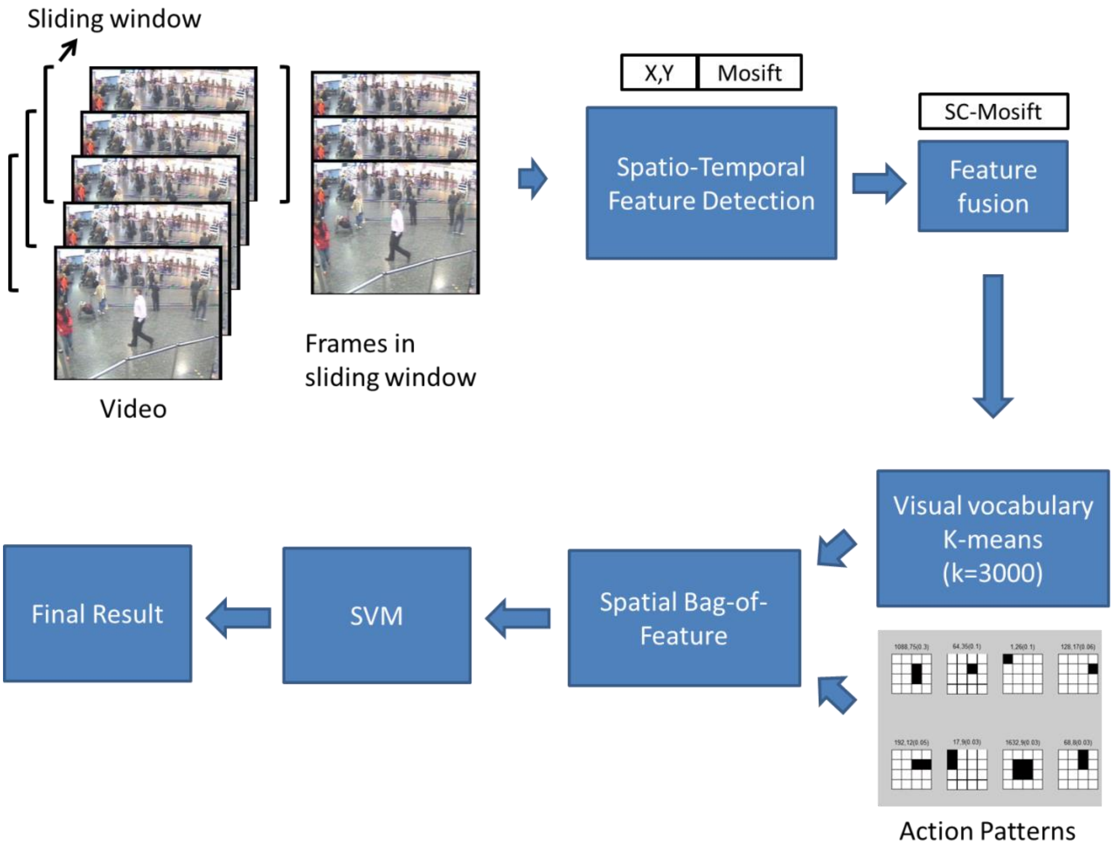


Figure 1 System Framework

In contrast, we extend the origin framework by two kinds of processing. Firstly, we replace MoSIFT with our SC-MoSIFT, in order to add spatial information to origin features. Since actions are related to locations, some actions always took place at one location, like embrace event for the camera 3 always took place in the middle of frame. The new feature will not only represent the local appearance and motion, but also represent the location that the interest point occurs. And an argument β is introduced to balance the representation ability of these two components. Secondly, at the stage of Spatial Bag-of-Feature (SBoF), we use an action oriented method to divide a frame into sub-regions, and the resulting Bag-of-Feature (BoF) feature are derived by concatenating the BoF feature captured in each sub-region. As mentioned in [4], the frame is divided into 3*3, 1*3, and each block is a sub-region. Instead, our action oriented method considers the blocks that a whole action goes through as a sub-region. And we choose the seven most frequency sub-region to form the final BoF. Since some actions like person run, the action may come across many blocks, intuitively, the statistic of visual words among all blocks can enhance the repeated feature caused by person run.

3. Annotation Selection

Considering that MoSIFT [1] feature is a combination of SIFT and optical flow, all the interest points have sufficient motion in the optical flow pyramid. Therefore, the clips with an unobvious Object Put movement are not ideal training samples. As

QMUL-ACTIVA's result [5] showed that, a selective of annotation will sufficiently decrease the rate of False Alarm.

In our experiment, we removed the annotations that have too much person occultation, unobvious movement and wrong annotation.

4. SC-MOSIFT

We use SC-MoSIFT as our low level feature. SC-MoSIFT is a combination of MoSIFT feature $F_i = \{f_1, f_2, \cdots, f_{256}\}$ at location $P_i = \{x, y\}$ and its normalized form $P'_i = \{x', y'\}$ $x', y' \in [0,1]$. We use scaling coefficient β to set the weight of location information when doing k-means (see in Figure 1), $P'_i(\beta) = \{\beta x', \beta y'\}$. As our experience on PUMP dataset, we would obtain the best result when location information and MoSIFT feature are at the same scale.

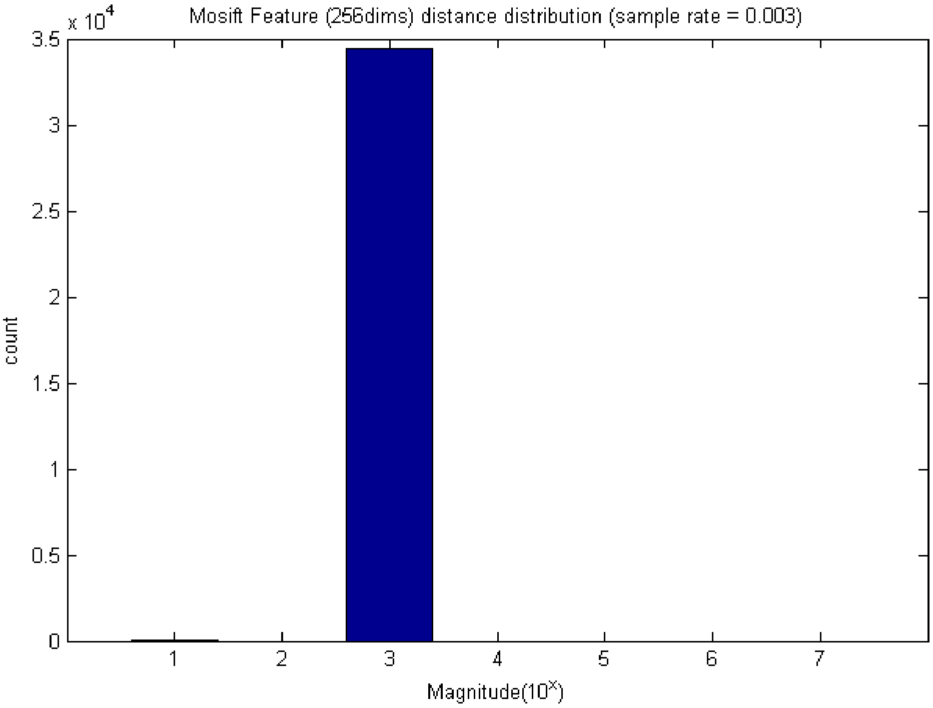
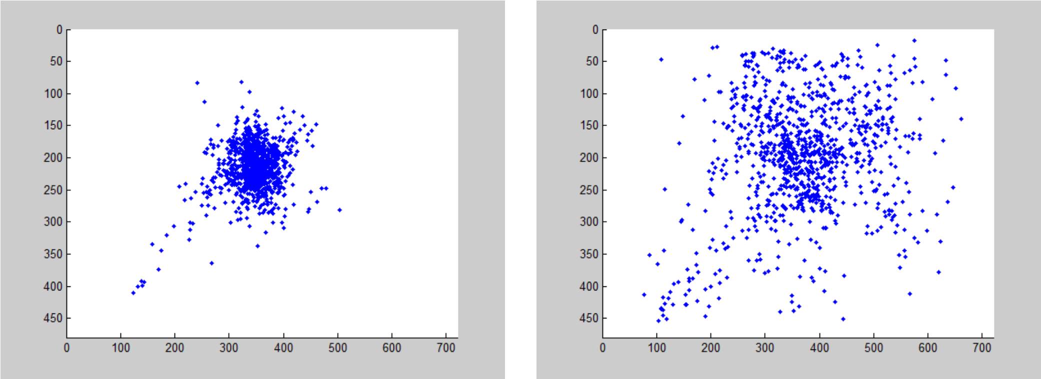


Figure 2 Distribution of the distance between MoSIFT feature

Refering to the distribution of the distance between two random MoSIFT features $d(P_i, P_j) = \|P_i - P_j\|$ in Figure 2, we set $\beta = 1000$.

We obtain 15% improvement in the measure of F1-score evaluated by leave-one-person-out cross-validation method on PUMP training data.

We can see the different visual words generated from the features with $\beta = 1$ (nearly the original MoSIFT feature) and $\beta = 1000$ (SC-MoSIFT). See Figure 3.



(a) (b)

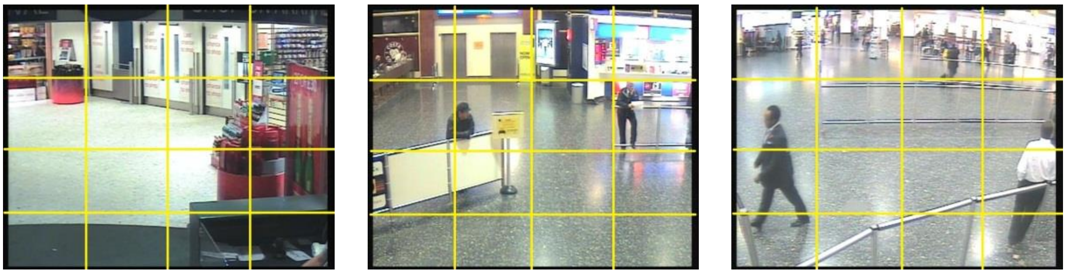
Figure 3 Distribution of visual words. 258 dims feature points draw on (x, y) dimensions

(a) $\beta=1$, (b) $\beta=1000$

Since actions are related to locations, some actions always took place at one location, like embrace event for the camera 3 always took place in the middle of frame. The new feature will not only represent the local appearance and motion, but also represent the location that the interest point occurs. This may result in an improvement on the performance.

5. Event Pattern

Event pattern illuminated the hot region of events where took place. We divided a frame in to 4*4 grids and showed in Figure 4.



(a) (b) (c)

Figure 4 Frame division. (a)Camera 1, (b) Camera 3, (c) Camera 5

We annotate an event with the continuous grids that it took place. And we call the continuous grids the pattern of an event. For example we label an Object Put event as showed in Figure 5, the man in the red area is throwing something into the dustbin. By counting the number of occurrences of each pattern, we can know how a whole event took place in a particular scene. See Figure 6.

As the people in CAM 2 are in quite small scale the movement is not sufficient, and the CAM 4 contains litter events, we do not label the actions in these two cameras and use the Run_1 method instead.

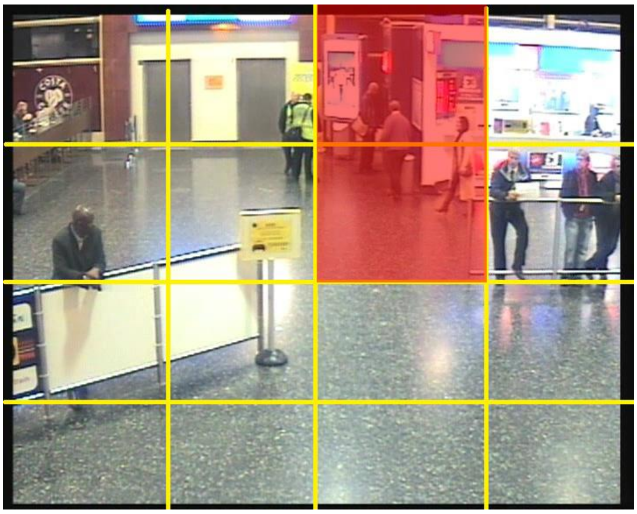
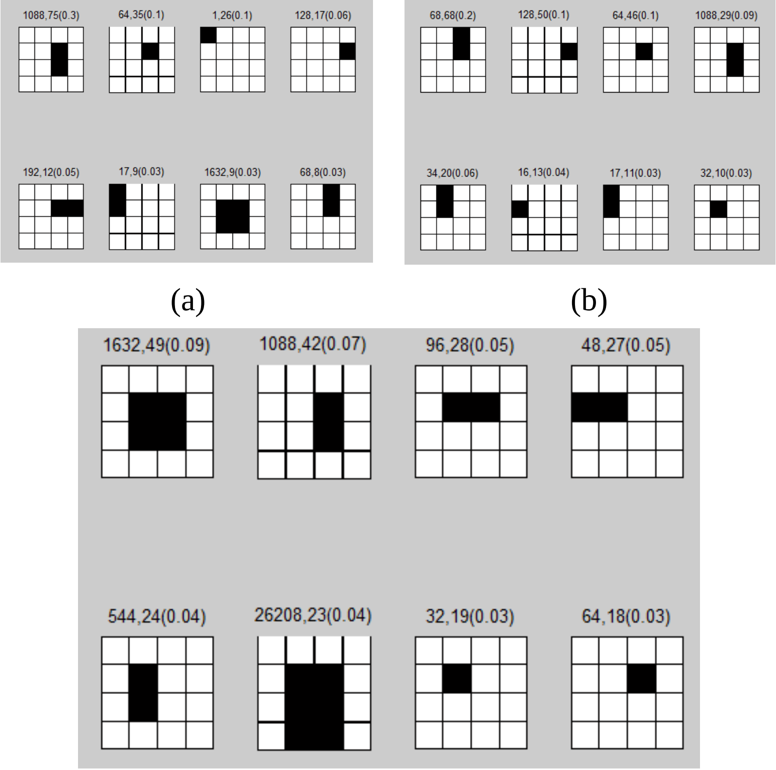


Figure 5 Object Put event pattern label



(c)

Figure 6 Object Put Pattern in different CAM, (a) CAM1, (b) CAM3, (c) CAM5. The figures show the eight most frequent patterns in each camera. The numbers above each square "BINARY, COUNT, (PERCENTILE)": BINARY is the binary representation of the pattern, COUNT is the appearance of the pattern, PERCENTILE is the rate of COUNT in all.

6. Event Pattern in Spatial BoF

In Run_2 we use the Event Pattern as the sub-region of spatial BoF, calculating BoF in each sub-region, and concatenating these BoF feature to the final Spatial BoF feature.

Comparing with the general division method [4] like 3*3, 1*3, the sub-regions determined by this method are able to contain a whole event in one sub-region. The statistic of visual words among all blocks that the action occurs can enhance the repeated feature caused by action. And this can help reduce noise interference, as the noise features are disordered.

7. Results and Future work

This is our first year's TRECVid attendance, we did not finished enough experiments on TRECVid dataset. With the result we already obtained, we can find that the new method we proposed got a comparable result to the former methods, see Table 2. In order to draw a convincing conclusion we need more experiments.

Table 2 Result

Run ID	Actual Decision DCR Analysis			MDCR
	#FA	#Miss	ADCR	
BIT_1	200	**609**	1.0463	1.0003
BIT_2	**127**	611	**1.0255**	**1.0000**

References

[1] M. Chen and A.Hauptmann, "MoSIFT: Reocgnizing Human Actions in Surveillance Videos," Carnegie Mellon University, 2009.

[2] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001.

[3] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton and G. Quenot, "TRECVID 2013 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics," in *Proceedings of TRECVID 2013*, NIST, USA, 2013.

[4] L. Zhang, L. Jiang, L. Bao, S. Takahashi, Y. Li and A. Hauptmann, "Informedia@TRECVID 2011: Surveillance Event Detection," 2011.

[5] F. Daniyal and A. Cavallaro, "QMUL-ACTIVA: 'Person Runs' detection for the TRECVID Surveillance Event Detection task," 2010.