TJUT-TJU@TRECVID 2011: Surveillance Event Detection

Zan GAO
Key Laboratory of Computer Vision and System (Tianjin University of Technology), Ministry of Education, Tianjin, 300384, P.R. China

An-An LIU, Yu-Ting SU, Zhong JI, Zhao-Xuan YANG
Multimedia Institute, *School of Electronic Information Engineering, Tianjin University, Tianjin, 300072, P.R. China*

Abstract: This year, we especially put our focus on analyzing motions in videos and the construction of hierarchical model. Firstly, we adopted a spatio-temporal interest point detector, which explicitly encodes appearance features together with motion information, to extract robust point features in a sliding window. And then the bag-of-word (BoW) approach is employed. After that, the hierarchical models are trained for each event and each camera. At the same time, we also discuss how to fuse results from different hierarchical models. Experiments show that the spatio-temporal feature is effective, and the hierarchical models are robust and stable, which are very helpful for improving our system's performance.

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, So MoSIFT algorithm[8] are proposed, which detects spatially distinctive interest points with substantial motions. 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. The experiments [8-12] showed that spatio-temporal feature was very robust, thus, in our experiments, we will adopt it.

In the following section, we will describe our system framework, and then the spatio-temporal interest point detector will be introduced. After that, the structure model is given. Finally, we will discuss and conclude the paper.

2. System Framework

Our team utilized the general framework to model human behavior with the philosophy of bag of

spatiotemporal feature (BoSTF) for individual event as shown in Figure 1. For each temporal sliding window in the video sequence, the spatio-temporal interest points are detected and formulated. Then the extracted spatio-temporal interest points are clustered into visual keywords and hierarchical SVM classifier is used for semantic event modeling. In the evaluation the video sequences are tested in the same way and the temporal adjacent events from different neighbor sliding windows are fused as identitical one. With these post-processing results the final decision will be given for the entire sequence.

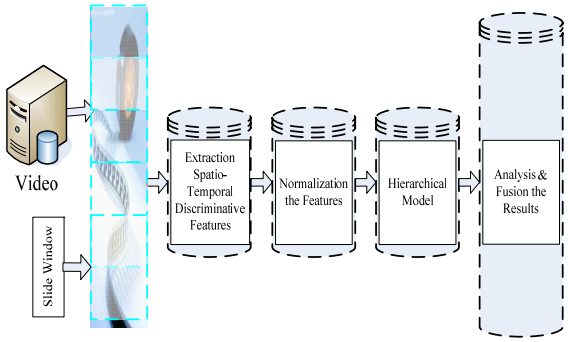

Figure.1 the framework of our surveillance event detection

3. Spatio-Temporal Interest Point Detector

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 spatio-temporal interest point detector 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

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 [14] 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.

3.2. Motion and Appearance Feature Description

After getting the spatio-temporal 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. Fig.2 shows the results of the spatio-temporal interest point detector in a Gatwick video key frame. It shows that the spatio-temporal features are able to clearly focus on areas with human activity.

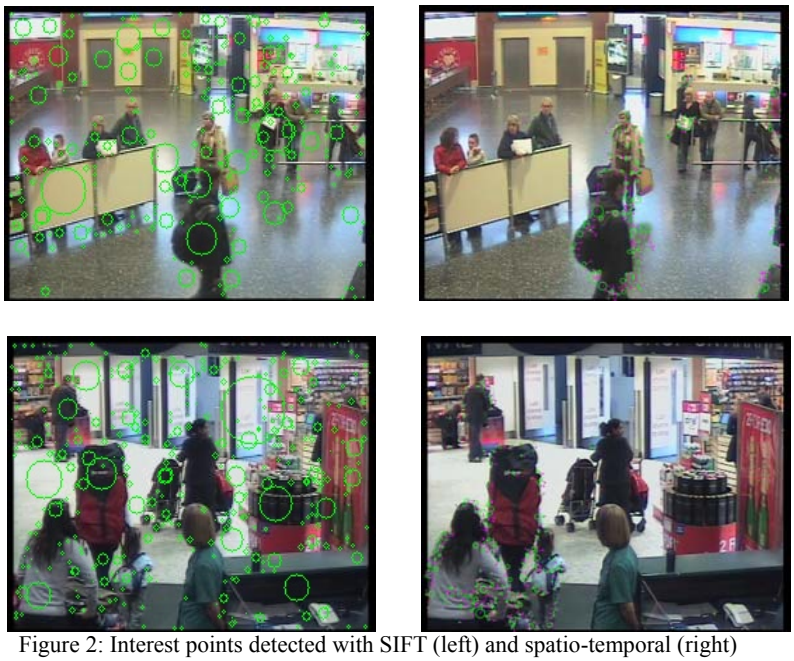

Figure 2: Interest points detected with SIFT (left) and spatio-temporal (right)

4. Hierarchical Model

In TRECVid 2011 Event Detection Evaluation [13], they provide 99 hours videos in the development set and about 44 hours videos in the evaluation set, where the videos were captured using 5 different cameras with image resolution 720×576 at 25 fps. In this dataset, the interesting events are very rare, and the sliding window is adopted. Thus, there are a lot of negative samples, and the model training will be very challenging. In order to solve the problem, the hierarchical models are trained. Firstly, we choose some negative samples randomly, and combine with all positive samples, and then the first model is trained on this dataset. Secondly, we will test all the negative samples in development dataset, and keep all the false positive samples. Thirdly, we will repeat the first step five times, and the last model will be kept as our classifier model. In the model training, the χ^2 kernel SVM [15] and one-against-all strategy are employed to construct action models.

5. Experiments and Discussion

In our experiment, the size of the slide window is set with experienced value 30 frames per second and the temporal step is the experienced value, 10 frames per second. The vocabulary size is 500 depending on our pervious experiments. For each sliding window, all of the spatio-temporal interest points in the window are projected into the vocabulary, and then spatiotemporal feature of the window can be represented by the histogram of the vocabulary. In total, there are 37 teams attending the task, but just Informedia and our team submit all the events evaluation. The performances of different cameras and different events are show in the following tables.

Table 1 Performance of our base algorithm under different thresholds

	Layer1					
	Threshold = 0.5			Threshold = 0.75		
Event	#CorDet	Act_DCR	Min_DCR	#CorDet	Act_DCR	Min_DCR
CellToEar	129	6.7798	0.9936	40	2.5059	0.9936
Embrace	146	7.6215	0.9764	137	3.6784	0.9984
ObjectPut	325	5.8081	0.9827	112	2.4691	0.9827
PeopleMeet	340	6.9715	0.9923	253	3.9086	0.9923
PeopleSplitUp	157	4.897	0.9677	60	2.5572	0.9847
PersonRuns	103	7.1725	0.9864	81	3.8417	0.9903
Pointing	763	7.6013	0.9951	335	2.7875	0.9951

Table 2 Performance of our hierarchical model under different thresholds

	Layer5					
	Threshold = 0.5			Threshold = 0.75		
Event	#CorDet	Act_DCR	Min_DCR	#CorDet	Act_DCR	Min_DCR
CellToEar	132	7.7288	0.9591	8	1.2368	**0.9591**
Embrace	137	7.8785	0.8575	91	1.5828	**0.8575**
ObjectPut	325	6.7706	0.8925	12	1.141	**0.8925**
PeopleMeet	298	5.9927	0.9041	104	1.4757	**0.9041**
PeopleSplitUp	161	5.0711	0.7449	11	1.0245	**0.7449**
PersonRuns	105	8.2658	0.8482	58	1.767	**0.8482**
Pointing	462	5.5653	0.9567	103	1.4153	**0.9567**

Table 3 Comparing our base model with hierarchical model in different cameras and the threshold is 0.75

	Camera1				*Camera2*			
	Layer1		Layer5		Layer1		Layer5	
Event	Act_DCR	Min_DCR	Act_DCR	Min_DCR	Act_DCR	Min_DCR	Act_DCR	Min_DCR
CellToEar	3.0266	0.9876	1.5837	0.9591	3.7202	0.9936	1.3599	0.8859
Embrace	2.9701	0.9959	1.7584	0.9681	3.7879	0.8863	0.9974	0.7592
ObjectPut	1.3617	0.9389	1.0016	0.8707	3.0323	0.9807	1.0459	0.8032
PeopleMeet	2.2688	0.9792	0.9983	0.7852	4.2624	0.9943	1.0165	0.7695
PeopleSplitUp	1	NaN	1	NaN	4.2366	0.9608	1.056	0.7818
PersonRuns	4.1849	0.9875	2.1697	0.975	5.4112	0.9279	1.7083	0.8368
Pointing	4.8851	0.9827	2.7031	0.9567	1.5787	0.9581	1.0033	0.7867

Table 4 Comparing our base model with hierarchical model in different cameras and the threshold is 0.75

	Camera3				*Camera5*			
	Layer1		Layer5		Layer1		Layer5	
Event	Act_DCR	Min_DCR	Act_DCR	Min_DCR	Act_DCR	Min_DCR	Act_DCR	Min_DCR
CellToEar	3.0325	0.9669	1.3178	0.8868	2.1765	0.9026	1.0098	0.766
Embrace	5.2553	0.9766	2.6341	0.922	6.9072	0.9984	2.6834	0.9279
ObjectPut	3.0912	0.9761	1.3442	0.8698	4.0198	0.9827	1.3217	0.8812
PeopleMeet	6.1841	0.9923	2.9128	0.9369	6.0846	0.9691	1.5803	0.8314
PeopleSplitUp	3.5918	0.9943	1.0803	0.9127	2.5302	0.961	1.041	0.8417
PersonRuns	4.1257	0.9728	1.8781	0.8499	5.3367	0.9903	2.5115	0.9298
Pointing	2.82	0.9839	1.205	0.9156	3.789	0.9947	1.1773	0.8564

From Table 1 and 2, we can see that the performances under different thresholds will be different, when the threshold is 0.5, its Actual DCR and Minimum DCR are much bigger than that under the threshold equals 0.75. Thus, finding the suitable threshold is very important. At the same time, we also can view that when hierarchical model is employed, the improvement is very large. For example, the actual DCR of 'CellToEar' under the base model and our hierarchical model are 2.5059 and 1.2368 respectively, and its minimum DCRs are 0.9936 and 0.9591 separately. Similarly, for other events, we also can get the same conclusions. Thus, the hierarchical model performance is much better than that of base model. Table

3 and 4 also demonstrate the performance comparing between base model and our hierarchical model under different cameras. Experiments show that under different cameras and different events, our hierarchical model still are much better than base model, and our model is very stable. Finally, we also consider how to fuse the results between our base model and hierarchical model, but we find that its improvement is very limit. In addition, the minimum DCRs in our hierarchical model for each event are under one, and the average minimum DCRs is 0.880429. When comparing with other teams, in total, our performance can reach the third place.

6. Acknowledgments

This work was supported in part by the Doctoral Fund of Ministry of Education of China (20090032110028), National Natural Science Foundation of China (61100124), Tianjin Research Program of Application Foundation and Advanced Technology (10JCYBJC25500), and 2011-2012 Innovation Foundation of Tianjin University.

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] Zan Gao, Ming-yu Chen, Alexander G. Hauptmann, Anni Cai, “Comparing Evaluation Protocols on the KTH Dataset”, Proc of ICPR2010 Workshop on Human Behavior Understanding (HBU2010), Istanbul, TURKEY Lecture Notes in Computer Science, v6219 LNCS, p88-100, 2010 (Oral Presentation, EI: 20104313332560).
[10] Zan Gao, Marcin Detyniecki, Ming-yu Chen, Alexander G. Hauptmann, Howard D. Wactlar, Anni Cai, “The Application of Spatio-temporal Feature and Multi-Sensor in Home Medical Devices”, International Journal of Digital Content Technology and its Applications,Vol. 4, No. 7, pp. 69 ~ 78, 2010.
[11] Huan Li, Lei Bao, Zan Gao, Arnold Overwijk, Wei Liu, Long-fei Zhang, Shoou-I Yu, Ming-yu Chen, Florian Metze and Alexander Hauptmann, Informedia @ TRECVID 2010 TRECVID Video Retrieval Evaluation Workshop, NIST, Gaitherburg, MD, November 2010.
[12] Ming-yu Chen, Huan Li, and Alexander Hauptmann. Informedia @ TRECVID 2009: Analyzing Video Motions.
[13] National Institute of Standards and Technology (NIST): TRECVID 2009 Evaluation for Surveillance Event Detection.http://www.nist.gov/speech/tests/trecvid/2010/ and http://www.itl.nist.gov/iad/mig/tests/trecvid/2009/doc/eventdet09-evalplan-v03.htm, 2009. 1, 7
[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.
[15] 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/.