SJTUBCMI at TRECVID 2012: Surveillance Event Detection

Keting Zhang, Wei Shi, Yangwei Wu, Liqing Zhang
Department of Computer Science and Engineering, Shanghai Jiaotong University
zzsnail@sjtu.edu.cn, bububut@gmail.com, wyw.wyw.good@163.com, zhang-lq@cs.sjtu.edu.cn

Abstract: In TRECVID 2012, our team takes part in the Surveillance Event Detection (SED) task and has finished four human events detection. We investigate an unsupervised learning approach based on an extended Independent Subspace Analysis model to extract spatio-temporal feature directly from the video data. The bag-of-words procedure and SVM classifier is used. We present the results and comparison of our primary run on the detection task.

1. Introduction

Human action recognition is a challenging problem in computer vision area and has attracted more and more interest in recent years. Action recognition and surveillance event detection have broad application in many areas especially public security applications such as airports, conference hall and supermarket. Nowadays, there are many effective algorithms which use hand designed local feature or spatio-temporal template. Laptev et al. [1] proposed space time interest point (STIP) which is an extension of spatio interest point into spatio-temporal domain. STIP is built on the Harris and Forstner interest point operators and reflects the significant variation in both space and time domain. MoSIFT[2] is another local feature descriptors used by CMU in SED task of TRECVID 2011 and a fairly good result is achieved. As a spatio-temporal template, Motion-Energy-Image (MEI) and Motion-History-Image (MHI)[3] template is proposed to match different kinds of actions. This spatio-temporal has many variations.

Recently, there is a growing interest in unsupervised feature learning directly from the raw data. Some methods are proposed such as sparse coding[4] and independent subspace analysis (ISA)[5]. In [5], the authors developed a two-layered convolutional ISA network for action recognition. They obtained the result better than the best published results until then. So we adopt this method as our main algorithm to finish the SED 2012 task.

In TRECVID 2012, our team takes part in the surveillance event detection task. In this task, the participants are provided about 150-hour surveillance video data recorded in London Gatwick airport. In this dataset, there are five cameras and hence five different real environments. Some of these environments are very complicated and consist of many moving people in different directions. So recognizing some kind of human action is a very challenging task. There are seven events to be detected which are “PersonRuns”, “CellToEar”, “ObjectPut”, “PeopleMeet”, “PeopleSplitUp”, “Embrace” and “Pointing”. Our team build a system to detect four human actions including “ObjectPut”, ”PeopleMeet”, ”PeopleSplitUp” and ”Pointing”.

The rest of this paper is organized as follows: Section 2 gives the description of our system framework. In section 3, the ISA algorithm adopted in our system is described briefly. Finally, we will present the experiment result and conclusion.

2. System framework

The framework of our surveillance events detection system is as shown in figure 1. Our system mainly includes four parts. Firstly, the video data is put into the system to extract the spatio-temporal feature after pre-processing. Then, visual words are got by K-means algorithm and bag-of-feature representation of video is generated. Finally, we use SVM with x^2 kernel [6] to classify the video sample.

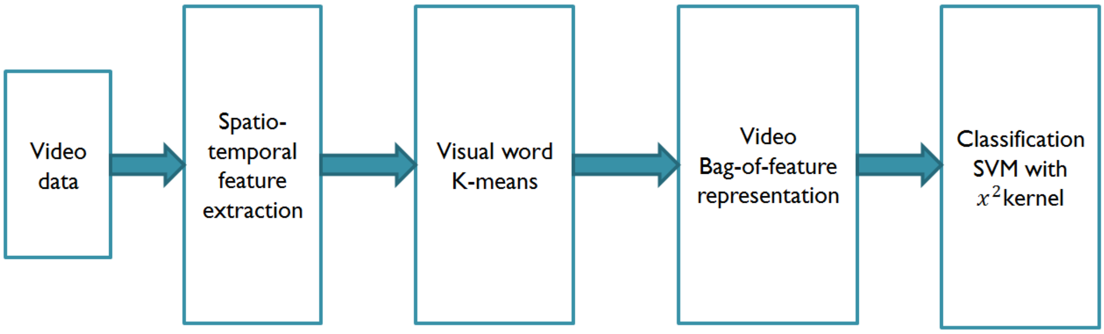
Video data
Spatio-temporal feature extraction
Visual word K-means
Video Bag-of-feature representation
Classification SVM with x^2kernel

Figure 1: The framework of our SED system

To represent the human action, we use the feature unsupervised learning directly from the video data based on ISA model. In [5], the authors proposed a two-layered stacked convolutional ISA network. After a preprocessing step with PCA, the whitening video data is fed into a ISA network. The standard ISA algorithm is very slow when the size of data is large. In this model, they make use of convolution and stacking to solve this problem. That is, they first train the network with small video patch, and then the output is used to convolve a larger region of the video. The generated result is then fed into the next layer. Similar to the first layer, before put into the next layer the data is also whitened and reduced dimensions using PCA. With this mechanism, the model enables the network to learn a hierarchical feature representation of human action. The first layer of this model can learn the feature representing a moving edge while the second one can learn to represent more complex shapes such as corners. To improve the recognition accuracy, the features learned from both of these two layers are combined for classification. For the details of this model, please refer to [5].

Because there are five different real-world environments in SED dataset, we train the network to get five different bases separately. During the training period, we find that some action duration in the development dataset is too short to form a complete video patch. So we abandoned these too short action slots. To get a video descriptor, we construct the visual words using K-means algorithm. We set K=3000. Then we train a classification model using SVM with x^2 kernel [6]. Following the idea in [7] and [8], we use the temporal sliding window method to get the video samples. The size of window is 60 frames. The sliding window step size we set is also 60 frames because of short of time to train. Obviously this reduces the classification accuracy.

In the SED task of next year, we will experiment different window size and different step size.

3. Experiment result

For the retrospective and interactive SED task, we submitted two primary run results for the events “ObjectPut”, ”PeopleMeet”, ”PeopleSplitUp” and ”Pointing” separately. The performance of the submission is evaluated by the actual Detection Cost Rate (DCR) and minimum DCR which is a linear combination of the two errors: Missed Detections (MD) and False Alarms (FA) [9]. DCR reflects a tradeoff between these two error types. And this is associated with the decision threshold to judge whether an action occurs or not. A lower DCR indicates the better performance. For the retrospective SED task, the performance of our system is summarized as shown in table 1.

From the table 1, it can be seen that our actual DCR is not very close to the min DCR which means the decision threshold we adopted is not very proper. This table also shows the comparison between our system with the best result in TRECVID 2012 SED task.

Table 1: Primary run results and comparison

Actions	Our Primary Run			Min DCR	Best Min DCR
	#FA	#Miss	Actual DCR		
ObjectPut	497	606	1.1388	1.0003	0.9983
PeopleMeet	989	374	1.1573	0.9956	0.9490
PeopleSplitUp	673	146	1.0014	0.9227	0.7882
Pointing	1179	1000	1.3274	1.0001	0.9770

4. Conclusion

In this paper, we give a brief description of our first system participating in the TRECVID 2012 SED task. We have finished four human events detection and submitted a primary run result. Our system is mainly based on a two-layered stacked convolutional ISA model. In the future, we plan to extend the current model to get the better representation of action and the parameters setup.

References

[1] I. Laptev. On space-time interest points. International Journal of Computer Vision, 64(2):107-123, 2005.

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

[3] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(3):257-267, 2001.

[4] B.A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583):607-609,1996.

[5] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal

features for action recognition with independent subspace analysis. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3361-3368. IEEE, 2011.

[6] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In BMVC 2009- British Machine Vision Conference, 2009.

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

[8] L. Zhang, L. Jiang, L. Bao, S. Takahashi, Y. Li and A. Hauptmann. Informedia@TRECVID 2011: Surveillance Event Detection
http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.cmu.sed.slides.pdf

[9] 2011 TRECVID Surveillance Event Detection (SED) Evaluation Plan
http://www.itl.nist.gov/iad/mig/tests/trecvid/2012/doc/SED12-EvalPlan-v03.html