A system for TRECVID MED by MCIS

Yang Feng, Wanchen Sui and Xinxiao Wu
Beijing Laboratory of Intelligent Information Technology, School of Computer Science
Beijing Institute of Technology, Beijing 100081, P.R. China
Email:{fengyangbit,suiwanchen,wuxinxiao}@bit.edu.cn

***Abstract*—We designed a simple system for the 2014 TRECVID Multimedia Event Detection [1]. Except the videos provided by NIST, we only used the BVLC Reference CaffeNet model file distributed besides Caffe [2]. Our system follows the standard pipeline and consists two parts: feature extraction and classification. The feature extraction part is implemented by Caffe and the classification is implemented by LIBSVM [3]. Based on the results, we think that the contribution mainly comes from the feature extraction part. We learned that Convolutional Neural Networks (CNN) is a powerfully model and hope that a easy accessible spatio-temporal CNN model for videos will be available soon.**

I. Introduction

Videos are ubiquitous nowadays. Automatically detecting events in these videos has become an important research topic due to its usefulness in video management, video retrieval and video surveillance. There are large intra-class variations in the unconstrained event videos, which makes it very challenging to recognize event in them.

Convolutional Neural Networks (CNN) [4] have achieved state-of-the-art performance in many computer vision problems. It is rarely used in video event recognition because training CNN usually requires huge amount of data and computation. Motivated by [5], we use a 2d CNN for video feature extraction to alleviate the computation cost. Each video is represented as the average of CNN features of the video frames. The CNN feature of a single frame is extracted by the ImageNet 2012 winning model. Although no temporal information is taken into account in the CNN feature, this video representation is powerful in event classification because the CNN feature can express the appearance information in the frames very well. The results demonstrated the video feature is effective in multimedia event detection.

II. Method

1) Metadata Generator: To extract the feature for a video, we first snip 100 frames in the video with equal step. The step is set as $\frac{total\ number\ of\ frames}{100}$. Then we feed the frame to the ImageNet 2012 and get the last fully connected layer as the frame feature. Finally, the video is represented by the average of the features of the video frames. With the existing software Caffe [2] or ConvNet, we can extract the CNN feature for videos easily.

2) Event Query Generator: We use the standard classifier LIBSVM [3] for classification. Because the positive training videos and background videos are heavily imbalanced, we use different soft margin parameter C for them. Specifically, we set $C = 32$ for the positive samples, $C = 3.2$ for the near miss videos and $C = 0.32$ for the background videos. We use the RBF kernel and keep the parameter g as the default value.

III. Conclusion

We have proposed an effective video feature for event detection. With Caffe, we can implement this feature easily and efficiently.

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot, “Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” in *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[2] Y. Jia, “Caffe: An open source convolutional architecture for fast feature embedding,” http://caffe.berkeleyvision.org/, 2013.

[3] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *CVPR*. IEEE, June 2014, pp. 1725–1732.