IIPWHU@TRECVID 2017
Surveillance Event Detection

Bin Xu, Weihang Liao, Wentao Bao, Daiqin Yang, Zhenzheng Chen*

School of Remote Sensing and Information Engineering, Wuhan University

Wuhan, China 430079

zzchen@whu.edu.cn

Abstract

In this paper, we present a system based on 3D convolutional neural network dealing with Surveillance Event Detection (SED) task in TRECVID 2017. In the proposed system, surveillance videos are decomposed to small fixed-sized video clips, which will be sent to CNN with 3D convolution layers to train models for classifying later. In the whole evaluation video, whether a clip contains required events or not depends on the clips classified through CNN model. The training data we used is part of Gatwick development data and we conduct the system on both the evaluation videos and the Group Dynamic Subset.

1. Introduction

The wide availability of low-cost sensors and processors has greatly promoted deployments of video surveillance systems [1]. Intelligent video surveillance applications such as automatic event detection have a significant impact on home security and public security. In the last few decades, most research of human action recognition mainly experiments on controlled environment with clear background where explicit actions are performed with limited actors. However, in real-world surveillance videos, due to challenges of large variances of viewpoint, scaling, lighting, cluttered background, it is almost impossible for us to have the ideal situation. Under the circumstances, the TRECVID [2] surveillance event detection (SED) task is provided to evaluate event detection in real-world surveillance settings. In TRECVID 2017 [3], the test data is the same data that was made available to participants for previous SED evaluations, which is about 100-hour surveillance videos under five camera views from the London Gatwick International Airport with annotations of event labels. Participates' systems should output detection results for events in the following list: PersonRuns, CellToEar, ObjectPut, PeopleMeet, PeopleSplitUp, Embrace, and Pointing. The main evaluation will be implemented using a 9-hour subset of the multi-camera airport surveillance domain evaluation data and a Group Dynamic Subset using only 2 hours of this video and limited to the Embrace, PeopleMeet and PeopleSplitUp events is also included this year.

We designed a system based on video feature extraction and 3D convolutional neural network dealing with surveillance event task in TRECVID 2017. The remainder of this paper is organized as follows. Section 2 introduce the overall retrospective system architecture, which contains preprocessing the original data, extracting feature form frame images and how the CNN works. The result of our approach performed on the TRECVID SED task is given in Section 3, and we conclude this paper in Section 4.

2. Retrospective System

2.1 Video Feature Extraction

Since videos in the dataset may contain different frames and the duration varies a lot. To extract features inside the spatial-temporal frames better, we first decompose each whole video to a set of small video clips. To reach the trade-off between efficiency and performance, each clip contains eight successive frames with four frames overlap. The famous architecture usually called C3D [4] is adopted for extracting features from each clip. It extends the traditional 2D convolutional to 3D. The architecture utilizes small $3 \times 3 \times 3$ convolution kernels in all 3D convolutional layers, which demonstrated useful in feature extraction.

2.2 3D Convolutional Neural Network

With the help of 3D convolutional layers. We can directly convert the event detection task to video clip classification problem. The video clips of parts of TRECVID 2008 dataset with available annotations will be used as training data for the 3D convolutional neural network. The number of CNN output are eight, corresponding to the seven required events and none required events, which is used as negative samples. The CNN architecture is shown in Fig. 1. The video clips of evaluation dataset are then sent to CNN to classify. The results of this step are the class of each clip with corresponding possibility value.

Thresholds of possibility value will be set to define whether the clips are considered as the corresponding event. Then the adjacent clips belonging to the same event will be combined as a consecutive event. The framework of our system can be seen in Fig. 2.

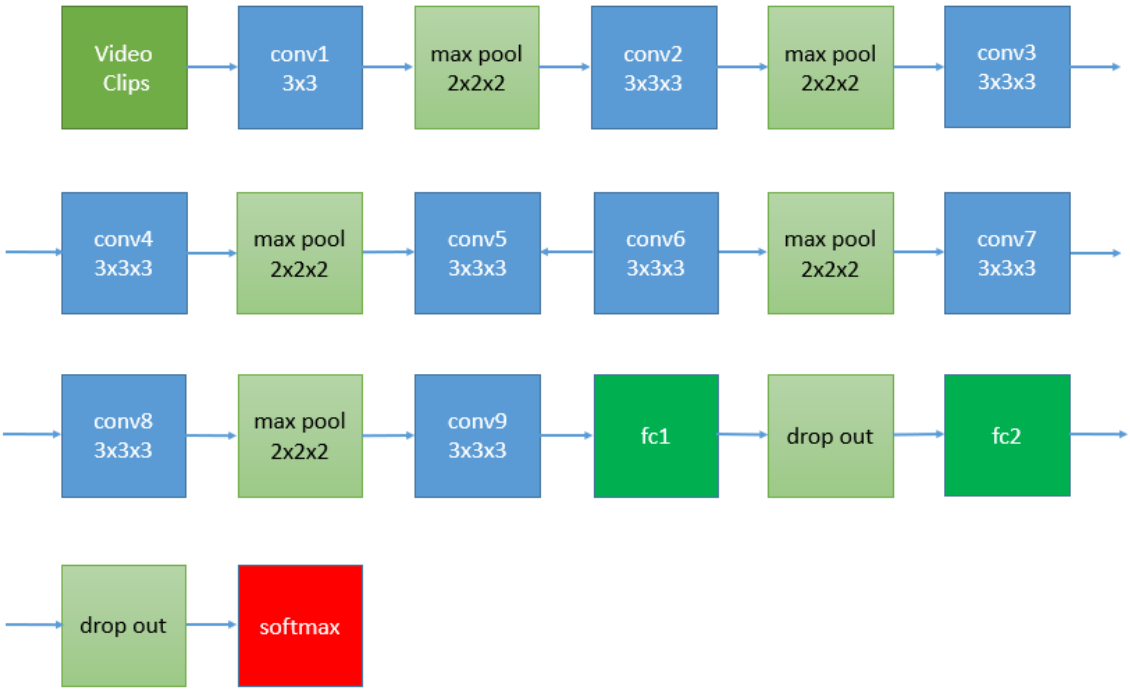

Fig. 1 The architecture of our network

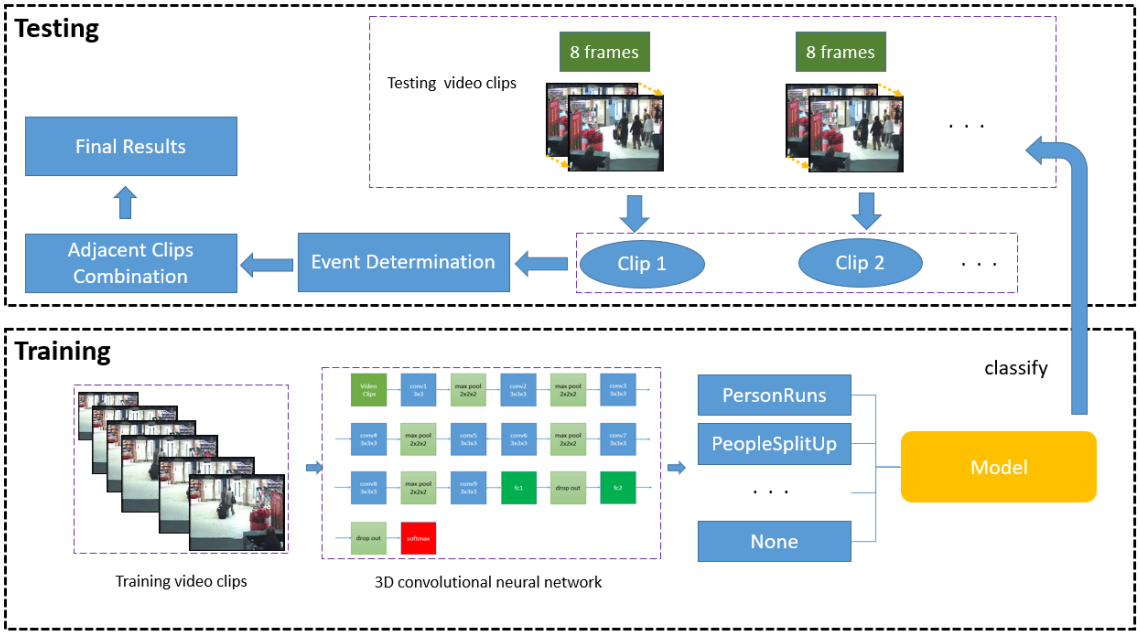

Fig. 2 The framework of system

3. Experiments

We applied our approach to the SED 2017 retrospective task. The CNN model of the system is trained on a DELL T5810 server that comprises 1 Intel Xeon E5 CPU,

32GB memory, and a single TITAN GTX1080 GPU. This year, the main evaluation (EVAL17) will be implemented using a subset of the multi-camera airport surveillance domain evaluation data collected by the Home Office Scientific Development Branch (HOSDB). A Group Dynamic Subset (SUB17) is limited to the Embrace, PeopleMeet and PeopleSplitUp events in 2017. Table 1 and table 2 shows our main evaluation and subset results provided by NIST. Our system only contains all the required events.

Table1: The actual DCR and minimum DCR of the EVAL17 result

Event	ActDCR	MinDCR
CellToEar	1.0549	1.0005
Embrace	1.2539	1.0005
ObjectPut	1.5944	1.0005
PeopleMeet	1.7796	1.0005
PeopleSplitUp	1.2829	1.0005
PersonRuns	1.7196	0.9921
Pointing	1.7900	0.9983

Table2: The actual DCR and minimum DCR of the SUB17 result

Event	ActDCR	MinDCR
Embrace	1.0429	0.9459
PeopleMeet	1.6962	1.0021
PeopleSplitUp	1.2802	0.9940

4. Conclusion

In this paper we have presented the detailed implementation of our system participated in TRECVID 2017. The system is applied to all seven events in EVAL 17 and all three events in SUB17. Video clips are generated from original dataset and used as the input for classification. CNN models are trained to deal with the evaluation videos. The current result is not good enough and more works need to be done to improve the performance in the future.

References

[1] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, pp. 3-19, 2013.

[2] A. F. Smeaton, P. Over and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, Santa Barbara, California, USA, 2006, pp. 321-330.

[3] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y.

Graham, W. Kraaij, G. Quénot, M. Eskevich, R. Ordelman, G. J. F. Jones and B. Huet, “TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking”, in *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, *ICCV* 2015.