IIPWHU@TRECVID 2016
Surveillance Event Detection

Zhongling Wang[1], **Jinghao Lu**[1], **Yisheng He**[1], **Siyuan Li**[1], **Bin Xu**[2], **Zhenzhong Chen**[1,2,†]
[1]*School of Computer Science, Wuhan University, Wuhan, China*
[2]*School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China*
zzchen@whu.edu.cn

Abstract

In this report we describe our system used in the TRECVID 2016[1] Surveillance Event Detection task. Our runid is *IIPWHU p-WhuIIPSubmission_2*. We proposed this system for detecting certain events in surveillance videos automatically. Our system is both trained and tested using 11 hour subset of the multi-camera airport surveillance domain evaluation data and the Group Dynamic Subset which contain 2 hours of video is also used. The videos with annotations are preprocessed using optical flow[2] to extract motion features, and then a model is obtained for later classification task by applying a convolutional neural network.

I. Introduction

With the increasing demand of surveillance video analysis, it is interminable and impossible for naked eye to detect abnormal events among enormous videos. So an efficient system is required for this task. We mainly focus on the events of PeopleMeet and PeopleSplitUp in this system. The videos with annotations are preprocessed using optical flow[2] to extract motion features, and then a model is obtained for later classification task by applying a convolutional neural network. Rest of this paper is organized as follows. In Section 2 we introduce the overall retrospective system architecture. The result of our approach performed on this task is given in Section 3. Finally, the paper ends with conclusions in Section 4. The framework of the entire system is illustrated in Fig. 1.

II. Retrospective System

A. Optical Flow Feature Extraction

Optical flow is well-performed in describing motion in videos since it contains both amplitude and direction information of moving objects. In the preprocessing phase, taking the processing speed into consideration, the videos are decomposed into frames after we resize the videos into a smaller size of 360×288. And then, the optical flow vector is computed between consecutive 2 frames using Lucas Kanade approach[3]. After this, we visualize the obtained optical flow vectors to color images using Munsell Color System[4]. In the color image, moving directions are represented by different hues while chromas indicate different moving speeds. Several examples of these images are illustrated in Fig. 2. These images will be used as features in later training and testing process.

B. Convolutional Neural Network

The optical flow images with available annotations will be used as training data for the convolutional neural network. We apply the BVLC AlexNet architecture in Caffe[5], which is a replication of the model described in[6]. We reset the output number to 3 in the network, corresponding to the events of PeopleMeet, PeopleSplitUp and non-required events. And the ratio of these events is $1 : 1 : 3$, respectively. The network architecture is shown in Fig. 3. In the phase of testing, the optical flow images generated in the same way as the training images are sent to the network for classification. And we get 3 possibilities for each frame. Each testing video is segmented to shots of consecutive 75 frames. In each shot, we add up the possibilities of the same category separately. The category with the highest summation of possibility

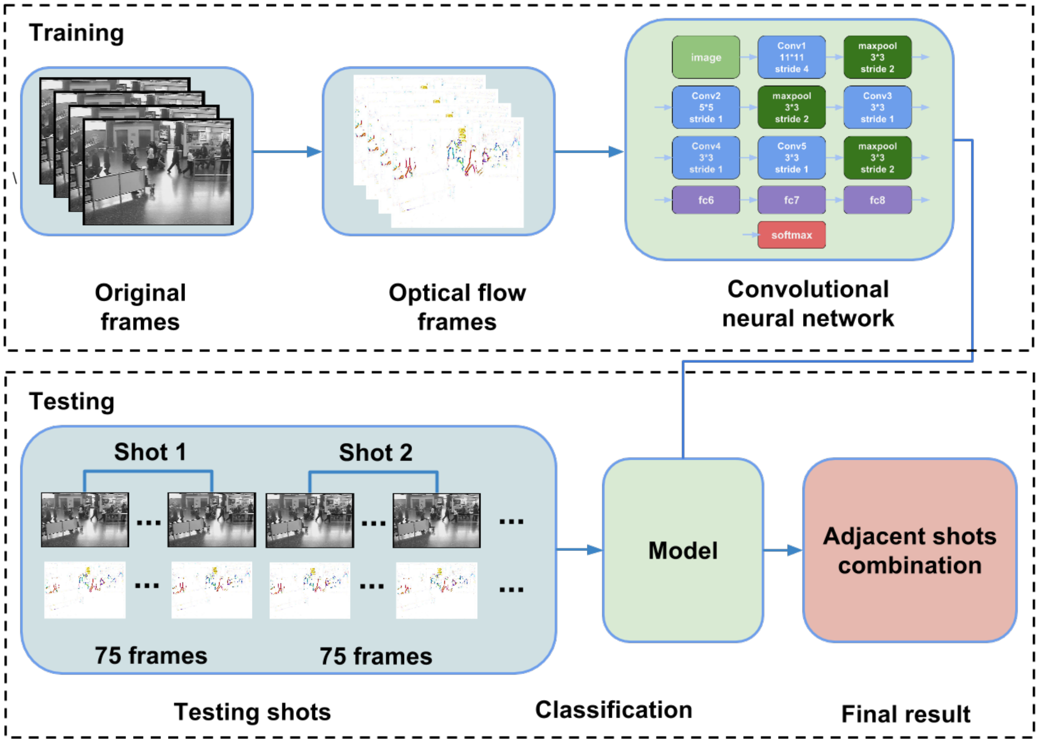

Fig. 1. The framework of our system

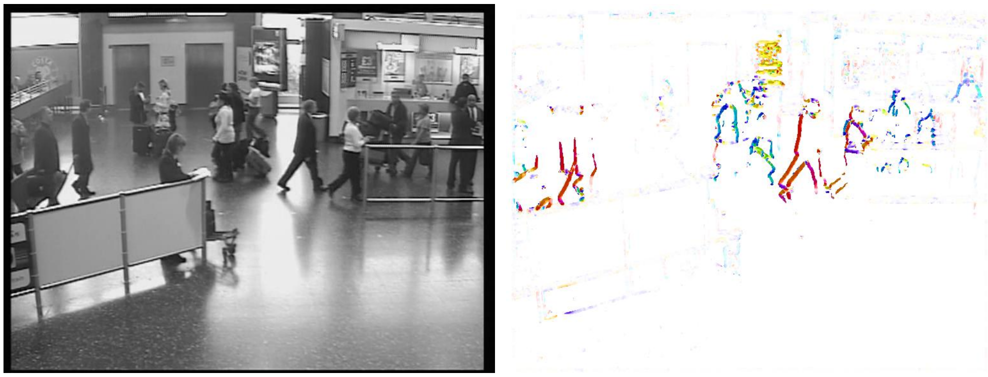

Fig. 2. Original image(left) and corresponding optical flow image(right)

is regarded as the category of this shot. At last, we set thresholds to determine whether if the event is detected in this shot. Then the adjacent shots having the same category will be combined into one long shot.

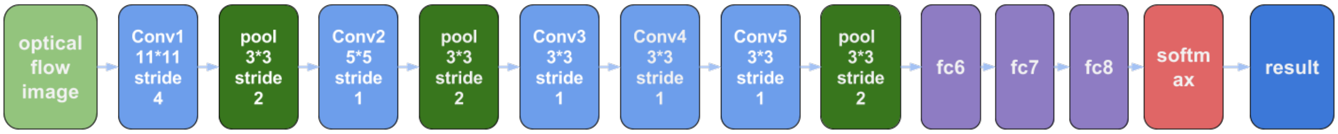

Fig. 3. The architecture of BVLC AlexNet

III. Experiments

Our system is both trained and tested using 11 hour subset of the multi-camera airport surveillance domain evaluation data collected by the Home Office Scientific Development Branch (HOSDB). The Group Dynamic Subset which contain 2 hours of video limited to the Embrace, PeopleMeet and PeopleSplitUp events is also used. Table I and Table II shows our EVAL16(IIPWHU p-WhuIIPSubmission_2) and SUB16(IIPWHU p-WhuIIPSubmission_2) results provided by NIST, respectively. Our system only contains events of PeopleMeet and PeopleSplitUp. The entire work is implemented on a work station with a 3.5GHz CPU, 16GB memory and a Quadro K2000 GPU.

TABLE I
EVAL16 result

Event	#CorDet	FA	#Miss	ActDCR
PeopleMeet	28	473	295	1.1495
PeopleSplitUp	32	711	144	1.1732

TABLE II
SUB16 result

Event	#CorDet	FA	#Miss	ActDCR
PeopleMeet	10	139	105	1.2104
PeopleSplitUp	22	186	75	1.1711

IV. Conclusion

In this paper we have presented the detailed implementation of our system participated in TRECVID SED 2016. The system is designed for events of PeopleMeet and PeopleSplitUp. Optical flow images are generated from original videos and used as features for classification. CNN models are trained to deal with the evaluation videos. However, the result is not good enough and more works need to be done to improve the performance.

References

[1] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Qunot, Maria Eskevich, Robin Aly, and Roeland Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[3] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.

[4] Albert Henry Munsell. *A color notation*. Munsell color company, 1919.

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.