HRI Team @ TRECVID 2016: Surveillance event detection

Ping Yang, Jiang Xiong, Di Xie, Shiliang Pu
Hikvision Research Institute, China

***Abstract*: Recent advances in object detection are driven by the success of deep neural networks, such as Faster R-CNN, SSD, etc. A system based on deep neural networks is designed for TRECVID 2016 Surveillance Events Detection. In this work, we utilize two neural networks: Region Proposal Network and Detection Neural Network. Region Proposal Network is trained end-to-end to generate high-quality region proposals, then Detection Neural Networks is used to determine whether the region proposals are target or not. Because surveillance events involve consecutive multi-frame, both of the two neural networks not only combine convolutional and recurrent neural networks, but also take multi-frame as input. Convolutional neural network extract the feature of frame, while recurrent neural network combine the feature of adjacent frames in video properly. In the training of two networks, in order to deal with the highly imbalanced nature of surveillance data, hard example mining algorithm is introduced, which make training more effective and efficient.**

Keyword: surveillance event detection, convolution neural network, recurrent neural network

I. INTRODUCTION

Surveillance event detection (SED), aiming at localizing and recognizing the temporal range of specific event, is a fundamental problem of variety of application for security and safety concerns. TRECVID sets the Surveillance Event Detection (SET) task to evaluate event detection in real-word surveillance settings. The primary condition for this task will be single-camera input. In TRECVID 2016, SED provides a corpus of 144-hour video under five camera views from the London Gatwick International Airport. In this dataset, about 100-hour videos can be used as development set with annotations of temporal extents and event labels. Our system is evaluated on three events, such as Personruns, Embrace, Pointing. In order to detect different events, our system should be adopted accordingly.

The rest of this paper is organized as follows. Section 2 introduces the system roughly, which include the structure of neural networks and post processing. Detailed descriptions for three events are provided in section 3. Section 4 present experiment result and discussions. Finally, we conclude in section 5.

II. OVERVIEW

Recent advances in object detection are driven by the success of deep learning, i.e. Fast R-CNN[1], Faster R-CNN[2], SSD[3], etc. In order to detect surveillance events, Faster R-CNN is applied in our system. However, Standard Faster R-CNN only confines on single image. In surveillance setting, only single frame is not enough to recognize some special event in most cases, which involve a series of different actions. In order to overcome this disadvantage, our system provides plenty of modification on Faster R-CNN, which include setting multi-frames as input and adding recurrent neural network. Comparing with single image, multiple sequential frames of surveillance video not only refer to more actions, but also contain plenty of temporal information. Recurrent neural network have emerged as an effective and scalable model for several learning problems related to sequential data, such as handwriting recognition and speech recognition. Our experiments demonstrate that the recurrent neural network also is efficient for events detection for sequential frames in surveillance video.

In our system, we pay attention to the events of Personruns, Embrace and Pointing. Neural network will be adapted to detection event accordingly. As demonstrated in Fig 1, Similarly as Faster R-CNN, our system consists of two neural networks: (1) Region Proposal Network (RPN), (2) detection network. An RPN is a fully convolution network that simultaneously predicts object bounds and object scores at each position. After training, the RPN is capable of generating high-quality region proposals. Then these region proposals are scored by detection network aiming at detecting required events. For simplification, RPN and detection network may be merged into a single network by sharing their convolutional feature, just as Faster R-CNN. At the same time, recurrent neural network is applied to fuse the feature of multi-frames or region proposals. According to different event detection, the number of input frame and position of recurrent neural network may be adapted.

Our system aims at locating and recognizing the target with special action accurately, so all positive frames with target are collected from TRECVID [6] as training samples and labeled with bound box, which contain the target properly, just as shown in Fig. 2. For the existence of recurrent neural network, all training samples must be listed in chronological order. During the training of our neural network, hard example mining algorithm [5] is introduced, which make training more effective and efficient.

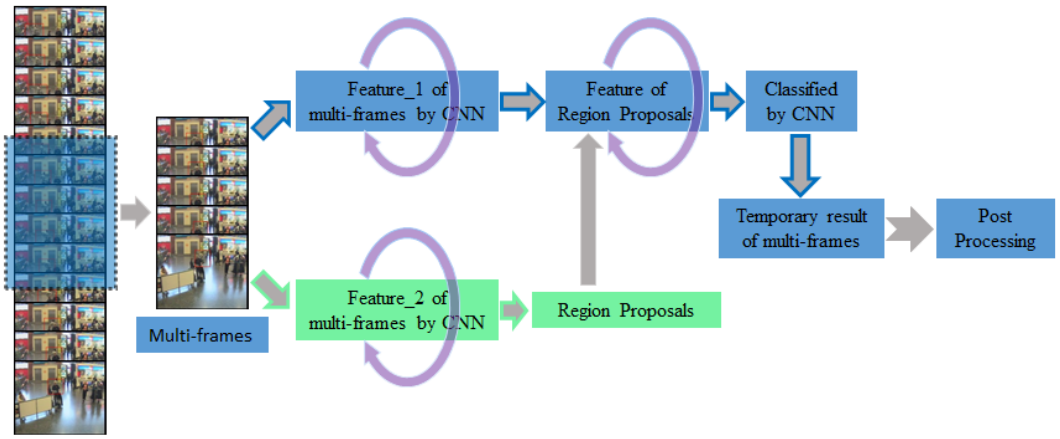


Fig. 1 An overview of our system for surveillance event detection.

(a) Embrace (b) Personruns (c) Pointing

Fig. 2 According to the events, positive frames is labeled with bound box.

After training, RPN takes multi-frames as input and output a set of rectangular object proposals with object score. Then, we refine the result from RPN by non-maximal suppression (NMS). Finally, detection neural network are designed to judge whether the region proposal with high score given by RPN is target or not.

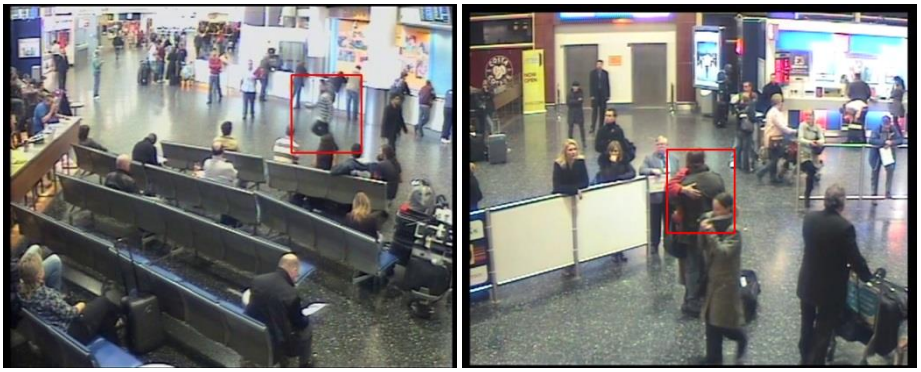
By the neural network, bound boxes of targets in sequent frames are obtained. Based on the position and size of bound boxes, trajectory of targets can be extracted. A trajectory result is shown in Fig. 3, where red line is the trajectory of center of bound boxes. Some properties of this trajectory are significant for judging whether the target is required target or not, such as speed and length.

Fig. 3 A series of rectangles with blue is the trajectory of targets, whose centers are shown by red.

III. EVENT DETECTION

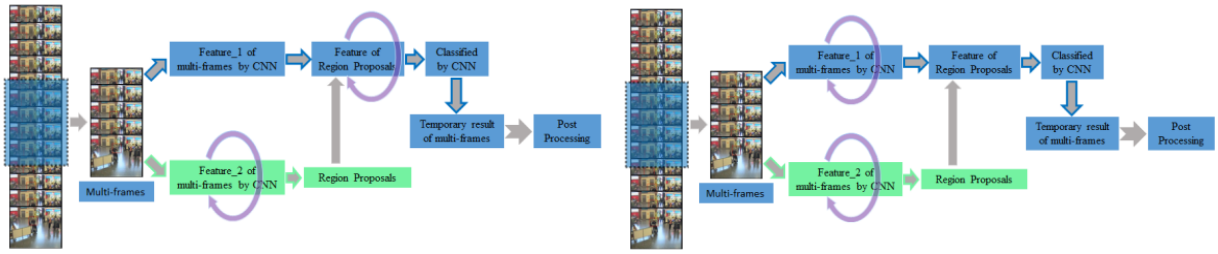
Our SED method mainly depends on the application of convolution neural network and recurrent neural network. At the same time, post-processing is used as aids.

Three events, such as Presonruns, Embrace and pointing, are evaluated in our system. These three events all involve series of poses. However, unlike the Pointing, Personruns and Embrace have more than one key-pose and can't be judged by only one frame in most cases.



(a) Personruns (b) Embrace

Fig. 4 One frame is difficult to localize and recognize the targets.



(a) Personruns (b) Embrace

Fig. 5 neural network for Personruns and Embrace.

A. PersonRun and Embrace

Just as shown in Fig4, one frame is difficult to localize and recognize the targets of Personruns or Embrace in most cases, because which may involve series of different actions or be disturbed by occlusion. So in our system, five adjacent frames of surveillance video are combined as input of neural network, and both two adjacent input have two copies.

At the same time, in order to take advantage of temporal information between adjacent frames, recurrent neural network is added to merge the feature of two adjacent inputs. Fig. 5 shows the neural network for Personruns and Embrace. The only difference of these two neural networks is the positon of recurrent neural networks in detection network. In detection network of Embrace, the recurrent neural network is designed to fuse the feature of all frames. While in detection network of Personruns, the recurrent neural network is used to feature combination of region proposals.

Based on the trajectory of targets, the speed can be estimated, which is significant for detection of Personruns. The higher speed of targets, the higher the probability of Personruns. Meanwhile, the speed also is useful for detection of Embrace. For Embrace, the speed should not be too high.

B. Pointing

The neural network for Pointing detection is demonstrated in Fig. 6. Different from Personruns and Embrace, Pointing only involves a key-pose, which result in that one frame is sufficient to detect it. So the single frame is set as the input of this network. Meanwhile, recurrent neural network is no longer necessary.

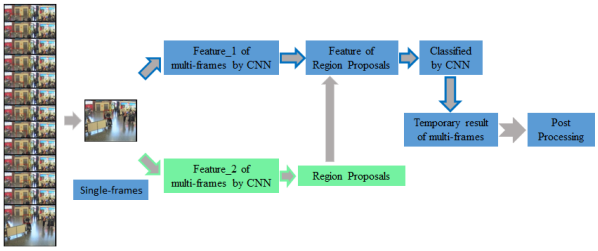


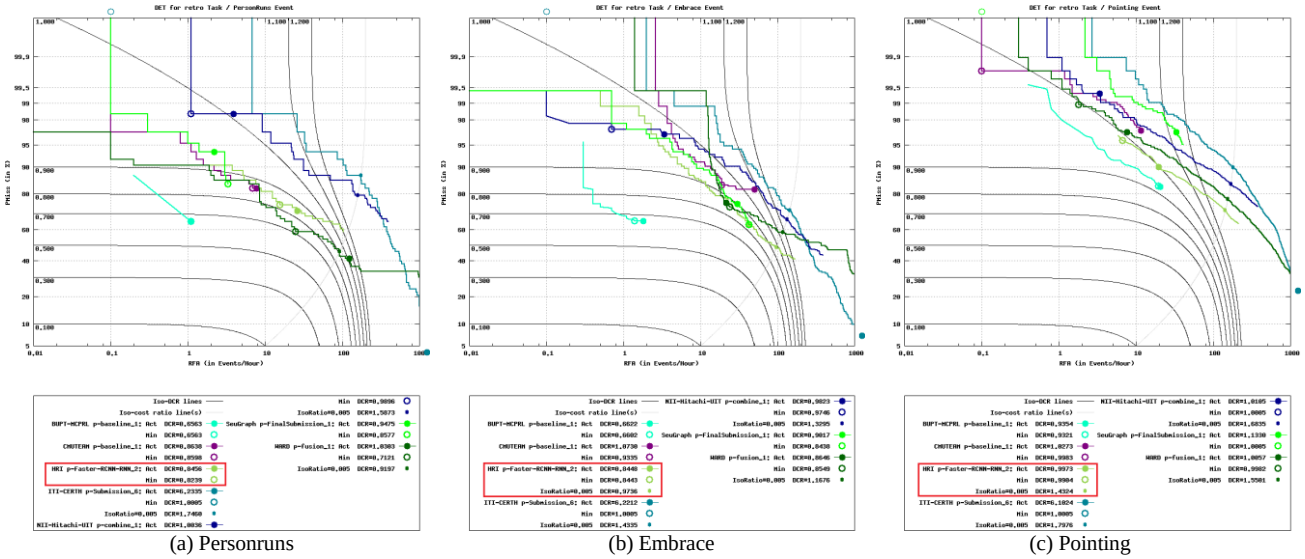
Fig. 6 neural network for Pointing.

IV. EXPERIMENTAL RESULTS

In TRECVID SET 2016 [6], EVALUATION Subset (EVA16) and the group Dynamic Subset (SUB16) are provided as the evaluation sets. Here, our system is only evaluated on EVA16 for detection of Personruns, Embrace and Pointing. The primary run results are shown in Tab 1. The Detection Error Tradeoff (DET) curves of event are shown in Fig.7.

Tab 1 Results of three events detection on EVA16.

Event	#CorDet	#FA	#Miss	ActDCR	MinDCR
Personruns	63	263	45	0.8456	0.8236
Embrace	173	430	109	0.8448	0.8443
Pointing	929	195	836	0.9973	0.9904



(a) Personruns (b) Embrace (c) Pointing

Fig. 7 the run result of our system for three events detection.

V. CONCLUSION

In this paper, we have presented detail implementation of our SED system participated in TRECVID 2016. The combination of convolution neural network and recurrent neural network provides an efficient network frame to detect surveillance events. Aided by the modified Faster R-CNN, our system is capable of locating and recognizing the required events exactly.

REFERENCES

[1] R Girshick. Fast r-cnn. Compute Science, 2015

[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016

[3] W Liu, D Anguelov, D Erhan, C Szegedy, S Reed. SSD: Single shot multiBox detector. Computer Science, 2015

[4] A Graves. Supervised sequence labelling with recurrent neural networks. Studies in Computational Intelligence, 2012

[5] A Shrivastava, A Gupta, R Girshick. Training region-based object detectors with online hard example mining. Conputer Vision and Pattern Recognition, 2016

[6] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, Roeland Ordelman. TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking. In Proceedings of TRECVID 2016. NIST, USA, 2016