한 국 정 보 화 진 흥 원

기술보고서

제목: Real-time Anomaly Detection using Deep Learning Based Network

과제명: KOREN SDI 기반 오픈플랫폼 실증

수탁기관 경희대학교 산학협력단

한 국 정 보 화 진 흥 원

- 목 차 -

1. Introduction

'Anomaly' in a general sense means something different from standard, normal or expected. This can be used to detect abnormal patterns in data, such as fraud detection and network intrusion, as well as to change into the age of big-data. It is also being used as an underlying technology to promote public safety by being applied to censorship of unusual, violent behavior or unnatural objects.

Surveillance videos should be able to capture a variety of realistic anomalies. Real world anomalous events are complicated and diverse. The goal of practical anomaly detection system is to timely signal an activity that deviates normal patterns and identify the time window of the occurring anomaly. Once an anomaly is detected, it can further be categorized into one of the specific activities using classification techniques. Some of the possible anomalies in pedestrian walkways are biker, skater, carts, etc. All those things can interfere with the pedestrian's course.

Figure 1 Examples of anomalies from UCSD Peds1[1] dataset

Current state-of-the-art object detection systems are variants of the following approach: hypothesize bounding boxes, resample pixels or features for each box, and apply a high-quality classifier. This pipeline has prevailed on detection benchmarks since the Selective Search work [2] through the current leading results on PASCAL VOC, COCO, and ILSVRC detection all based on Faster R-CNN [3] albeit with deeper features such as [4]. While accurate, these approaches have

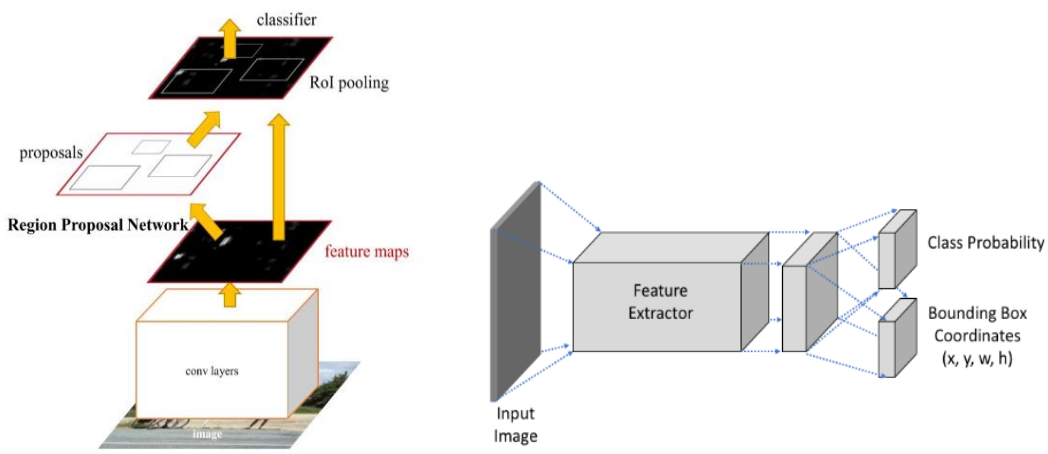
been too computationally intensive for embedded systems and, even with high-end hardware, too slow for real-time applications.

. Particularly, this study adopts the deep network based object detector i.e. single shot multibox detector (SSD) [5]. It does not resample pixels or features for bounding box hypotheses and is as accurate as approaches that do. This results in a significant improvement in speed for high-accuracy detection

The remainder of content is organized as follows. We describe deep learning based object detection model along with CCTV surveillance dataset preprocessing in Section 2. In Section 3, we present our total process of anomaly detection in real world CCTV surveillance video. Furthermore, we evaluate the system performance with testbed implementation in Section 4. Finally, we conclude this paper in Section 5.

2. Deep learning based Model and Preprocessings

2.1. Deep learning Model



(a) Two stage detector (b) One stage detector

Figure 1. Two mainstreams of deep learning based detector

Figure 1 depicts the two mainstreams of the detectors that use deep learning method:

(a) **Two-stage Detectors**: Detectors who consist of two stage to make detection result. The first stage decides the set of candidate proposals that could contain

objects and filters out the other regions. After that, the second stage classifies the proposals into each candidate classes and background. R-CNN [9] upgraded the second-stage classifier with using convolutional network, with large performance gains. However, the Faster RCNN framework [10] that the Region Proposal Networks (RPN) integrates proposal generation with the second-stage classifier into a single convolution network is the representation of this kind of detectors

(b) **One-stage Detectors**: Detectors who consist of just one stage to make detection result. They usually faster than two stage detectors since they perform feature extraction and boundary box detection at once. Recently SSD [5] and YOLO [11] are the representatives of this method.

2.2. Single Shot Multi-box Detection Network

The single shot multi-box detector (SSD), published in 2016, approach is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections. The early network layers are based on a standard architecture used for high quality image classification (truncated before any classification layers). Figure 2 shows the SSD network architecture. SSD has following key features

- **Multi-scale feature maps for detection:** convolutional feature layers are added to the end of the truncated base network. These layers decrease in size progressively and allow predictions of detections at multiple scales.
- **Convolutional predictors for detection:** each added feature layer (or optionally an existing feature layer from the base network) can produce a fixed set of detection predictions using a set of convolutional filters.
- **Default boxes and aspect ratios:** a set of default bounding boxes is associated with each feature map cell, for multiple feature maps at the top of the network. The default

boxes tile the feature map in a convolutional manner, so that the position of each box relative to its corresponding cell is fixed.

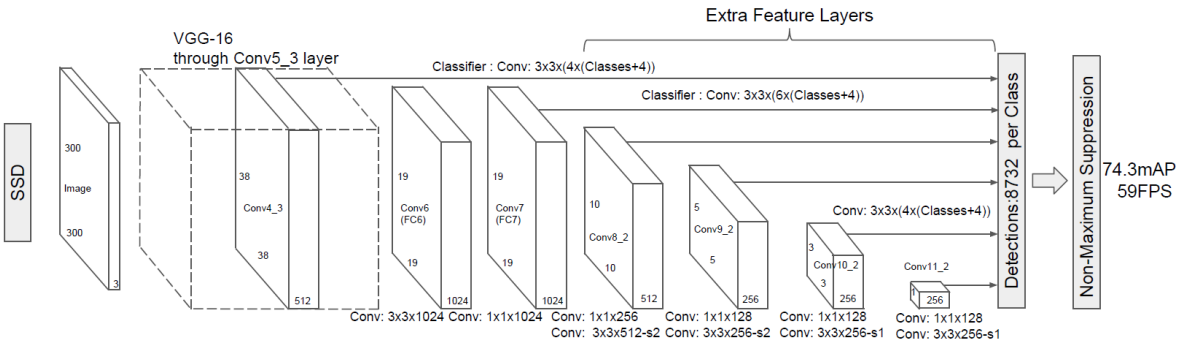


Figure 2 SSD network architecture

2.2-1 Training SSD

The way to train SSD network is derived from multi-box objective [6, 7] but not for the binary classification, but for handling multiple object categories. Overall function is a weighted sum of the localization loss (L_{loc}) that scores the accurate location of detected bounding boxes through the network, and the confidence loss (L_{conf}) that evaluates the reliability of class prediction result of the network:

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \quad (1)$$

Where N is number of matched default boxes. If N = 0 loss is set to 0. The localization loss is a Smooth L1 loss [6] between the predicted box (l) and the ground truth box (g) parameters. The class score is calculated with softmax loss over multi-class confidences. Finally, the parameter α is set to 1 determined by cross validation.

2.3. Constitute of Dataset

Figure 3. UCSD dataset and real-world CCTV dataset.

Figure 3 presents some examples of UCSD anomaly dataset and the real-world CCTV video dataset that are manually collected from JeonNam University.

UCSD Anomaly Detection Dataset defines an abnormal object when an object other than a pedestrian appears on the pedestrian road. In the data set, cart, wheelchair, skater, and biker are made up of abnormal objects. Therefore, training was performed on the data set to capture when an abnormal object appeared in a similar CCTV image.

However, the dataset is not labeled for all files, but for some files there is a zero-one label image. This is unsuitable for learning SSD. Therefore, we preceded the annotation work of transforming the existing label into the .txt file that displays the labels' names (cart, bicycle, skater, wheelchair) and coordinate values of the upper left and lower right of the label bounding box.

As can be seen from Figure 3, the UCSD dataset and the actual CCTV image have different characteristics such as the congestion of the pedestrian road and the color of the image, etc. Therefore, the performance of the SSD network is improved by using the actual CCTV image as the training set to construct the anomaly dataset. However, since additional labeling work is

required for the CCTV image, which is manually collected one, the labeling work was performed on both UCSD and actual CCTV data using the VGG image annotator tool [8].

3. Proposed Anomaly Detection Process

In this section, we present the whole process of our anomaly detection network. One of the most important point of this procedure is that it takes around 6 minutes to deal with those whole process. The overall automatic process consists of 5 stages.

1) Download the KOREN CCTV dataset from server
2) Cut down the CCTV dataset to consecutive image with 8.5 frame per sec.
3) All frames were taken to SSD network to detect the anomalies in each frame
4) Synthesize the detected frames into a format of mp4 to make video.
5) Send back the result video to server so that surveillance administrator can see the result

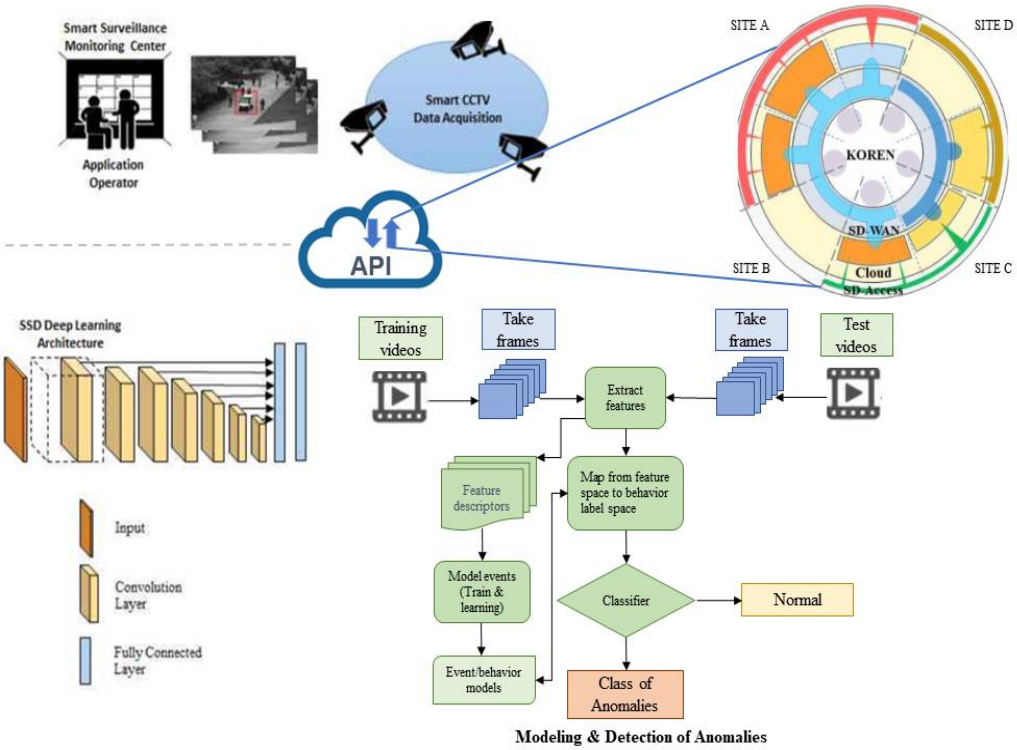


Figure 4 Overall architecture of deep learning based anomaly detection process

4. Experimental Result

For performance evaluation, we implement the proposed process to UCSD dataset and also applied them to real-world CCTV video dataset. In case of UCSD dataset, we just evaluate the performance of SSD network to observe that the network was trained properly or not. And for CCTV dataset, they go through the whole proposed process that introduced in Section 3, and show the performance of SSD network toward the real CCTV surveillance dataset

Considering the large number of boxes generated from SSD method, it is essential to perform non-maximum suppression (nms) efficiently during inference. By using a confidence threshold of 0.01, most boxes can be filter out. Then nms is applied with jaccard overlap of 0.45 per class and keep the top 10 detections per image. This step costs about 1.7 msec per image for SSD300 and 5 classes. we have achieved best accuracy 68% MAP at 0.5 threshold for UCSD dataset and best accuracy 75% MAP at 0.8 threshold on CCTV video dataset. Figure 3 shows the detection results from both datasets.

Figure 5 Detected anomalies from UCSD and CCTV videos

Figure 5 presents detected anomalies from UCSD and CCTV videos. From observing the upper left and lower right results, small occlusion cannot disturb the detection process of SSD.

5. Conclusion

In this study, we proposed a procedure of real-time CCTV surveillance network based on deep learning. Since SSD utilizes the multi-scale feature maps for detection, it could deal with the dynamic scale range of object event though the objects are small. Furthermore, we implemented some annotation process to real world CCTV video, to make more accurate anomaly detection system. As a result, we got approximately 65 fps from SSD, which can be considered to operate in real-time, and observed 75% MAP at 0.8 confidence threshold on CCTV videos and 65% MAP on UCSD dataset with 0.5 confidence threshold.

Reference

[1] http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

[2] Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV (2013)

[3] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. (2015)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)

[5] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. In: Springer, Cham (2016)

[6] Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: CVPR. (2014)

[7] Szegedy, C., Reed, S., Erhan, D., Anguelov, D.: Scalable, high-quality object detection. In: arXiv:1412.1441 v3 (2015)

[8] Dutta, A. and Gupta, A. and Zissermann, A.: Development and maintenance of VGG Image Annotator (VIA) is supported by EPSRC program grant Seebibyte: Visual Search for the Era of Big Data (EP/M013774/1) (2016)

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," In: NIPS (2015)

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," In: CVPR (2016)