AVS-Net: Automatic Visual Surveillance using Relation Network

Sein Jang,[1] **Young-Ho Park,**[2] **Aziz Nasridinov**[1]
[1]Chungbuk National University, Cheongju, 28644, South Korea, +82-10-7482-7227
{sein, aziz}@cbnu.ac.kr
[2]Sookmyung Women's University, Seoul, 04310, South Korea
yhpark@sm.ac.kr

Abstract

Visual surveillance through closed circuit television (CCTV) can help to prevent crime. In this paper, we propose an automatic visual surveillance network (AVS-Net), which simultaneously performs image processing and object detection to determine the dangers of situations captured by CCTV. In addition, we add a relation module to infer the relationships of the objects in the images. Experimental results show that the relation module greatly improves classification accuracy, even if there is not enough information.

Introduction

Crime is one of the most serious problems in modern society. To prevent crime, one efficient way is automatic visual surveillance through CCTV. There is not enough manpower to monitor every area through human operators. Meanwhile, due to the rapid development of deep learning, computer vision has shown remarkable progress in image classification and object detection. For this reason, there have been a variety of approaches to automatic visual surveillance using deep learning. These approaches can be divided into two categories: image classification and object detection.

Image classification approaches determines suspicious activities using feature extraction and classification techniques. For example, (Grega et al. 2016) used visual descriptors from the MPEG-7 feature extraction library for feature extraction, and used a support vector machine (SVM) to determine, and alert suspicious activities captured by CCTV. On the other hand, object detection approaches performs visual surveillance by identifying a dangerous object or behavior in the CCTV. (Nair, Mathew Gillroy, and Davies 2018) proposed i-SURVEILLANCE, which detects abnormal behavior by a person and alerts a human operator. The Tensorflow Object Detection API used to determine suspicious activities by detecting the persons on CCTV who exhibit abnormal behaviour. However, both approaches rely heavily on the performance of each visual surveillance technique.

In this paper, we propose automatic visual surveillance using a relation network (abbreviated as AVS-Net), which simultaneously performs image classification and object detection. The advantage of the proposed approach is that the two processes complement each other in determining suspicious activity. For example, in object detection approaches, the failure to detect a dangerous object can lead to failure of the visual surveillance. In contrast, under the proposed approach, even if the object detection fails, it can still perform visual surveillance through image classification. Furthermore, in order to increase the accuracy of visual surveillance, we apply a relation module. The relation module infers relationships between objects in the image and obtains detailed insights from the image, such as identifying different situations based on the location of each object and the distance between them.

Proposed Model

Figure 1 shows the overall flow of the proposed approach.

Image Classification. We used a convolutional neural network (CNN) to obtain a set of objects from an image. The 128 x 128 size input images pass through four convolution layers. The output of the CNN is a set of objects from feature maps of the image "i-objects", $I = (i_1, i_2, ..., i_n)$. Each pair of i-objects is treated as input to the relation module.

Object Detection. We used the you only look once (YOLO) (Redmon and Farhadi 2017), which is a model to perform object detection. We defined dangerous tools and actions to detect suspicious activity. Using a trained YOLO model to detect defined objects, we obtain a set of objects detected by YOLO "d-objects", $D = (d_1, d_2, ..., d_m)$. Each d-object is treated as input to the relation module.

Relational Reasoning. For relational reasoning between the image and the objects, we used the relation module (Santoro et al. 2017). Each i-object from the output of the CNN, $i_i \in \mathbb{R}$, is the i^{th} object that could comprise information about the background, a particular physical object, a texture, etc. The d-object, $d_k \in \mathbb{R}$, is the k^{th} object that could comprise the type, coordinates in the image, and the size of the detected object. In the relation module, the model learns the relations between objects in the image, so that our model can figure out the current situation in the image. As shown in Figure 1 (c), to compose the input for the relation module, we generate a tuple of objects by combining a pair of i-objects and a d-object as (i_i, i_j, d_k). Next, the object pairs passes through a multi-layer perceptron (MLP), where the model learns the relation between each object. We call the output of the first MLP a "relation feature". Finally, an

element-wise sum for all relation features is performed, and the result passes through the another round of MLP for classification.

$$R(I, D) = f(\sum_{i,j,k} r(i_i, i_j, d_k)), \quad (1)$$

Equation (1) is a composite function of the relation module, where f, r represents the four-layer MLP.

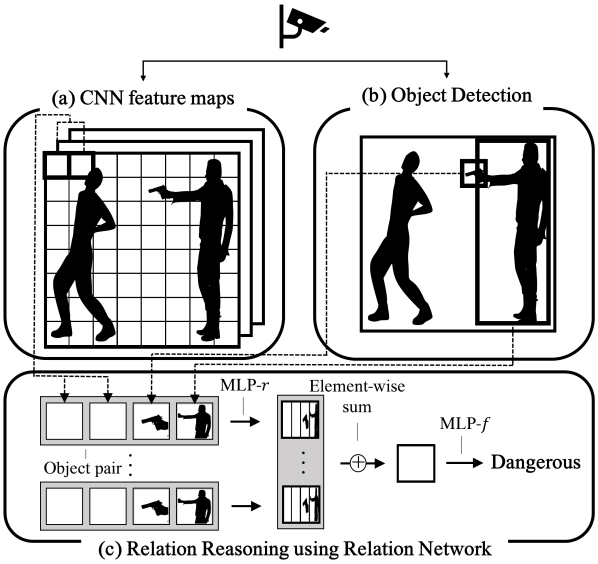

Figure 1: The overall flow of the proposed approach

Experiments and Results

The experimental results are described in detail below.

Dataset. For the experiment, we constructed a dataset that contains four situations (i.e., classes): normal, danger tool, potential crime, and dangerous. We shot videos of each situation, and generated frame images from the videos. A total of eight scenes contained four locations and two actors.

Baseline. We compared our model against several baseline methods, such as SVM, decision tree, random forest and gradient boosting. These baseline methods were implemented using the Scikit-Learn library.

Evaluation. We evaluated the model by measuring the prediction accuracy for the evaluation data set. In addition, as our hypothesis was that the model can perform automatic visual surveillance even with low performance from object detection, we made comparisons by adjusting object detection accuracy from 70% to 100%. To do that, we removed some of the object detection values randomly.

Results. Our proposed model outperforms the baseline methods (see Table 1). Even if the object detection has low performance, the accuracy of our model is 2% to 17% higher than the baseline methods. The reason for this result is that our model performs image processing and object detection simultaneously. Therefore, even an image with low resolution, or that fails under object detection, still has two processes that supplement each other. Moreover, since the proposed approach performs relational reasoning up to the last step, the model can figure out more details about the situation. A central contribution of this work is that the proposed approach can solve a problem that relies highly on object detection performance, which object detection approaches to automatic visual surveillance are struggling with the most.

Approach	Object Detection Accuracy			
	100%	90%	80%	70%
SVM	31%	27%	25%	23%
Decision-Tree	95%	89%	81%	70%
Random Forest	96%	88%	80%	71%
Gradient Boosting	96%	89%	80%	71%
MLP	32%	-	-	-
CNN+Object Detection	32%	-	-	-
Proposed Method	**96%**	**91%**	**90%**	**88%**

Table 1: Comparison of various baseline methods

Discussion and Conclusions

We propose AVS-Net, a model for automatic visual surveillance. Our model combines two different models: a CNN and object detection. A relation module is used to improve performance. Experiment results showed that (1) the proposed approach compensates for each process's deficiencies and shows high accuracy, compared to baseline methods; and (2) the proposed approach has the potential to handle all kinds of abnormal situations by extended data (i.e., Situations). For future work, we will generate a synthetic data set for more experiments, and furthermore, it would be interesting to experiment with real CCTV data in order to apply our model in real-life situations.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2016-0-00406, SIAT CCTV Cloud Platform).

References

Grega, M.; Matiolański, A.; Guzik, P.; and Leszczuk, M. 2016. Automated detection of firearms and knives in a cctv image. *Sensors* 16(1):47.

Nair, M.; Mathew Gillroy, N. J.; and Davies, J. 2018. i-surveillance crime monitoring and prevention using neural networks. *IRJET* 5(3):1231–1236.

Redmon, J., and Farhadi, A. 2017. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517–6525. IEEE.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, 4967–4976.