Motion-Attentive Network for Detecting Abnormal Situations in Surveillance Video

U-Ju Gim
Dept. of Computer Science
Chungbuk National Univ.
Cheongju, South Korea
kwj1217@cbnu.ac.kr

Jeong-Hun Kim
Dept. of Computer Science
Chungbuk National Univ.
Cheongju, South Korea
etyanue@cbnu.ac.kr

Kwan-Hee Yoo
Dept. of Computer Science
Chungbuk National Univ.
Cheongju, South Korea
khyoo@cbnu.ac.kr

Aziz Nasridinov*
Dept. of Computer Science
Chungbuk National Univ.
Cheongju, South Korea
aziz@cbnu.ac.kr

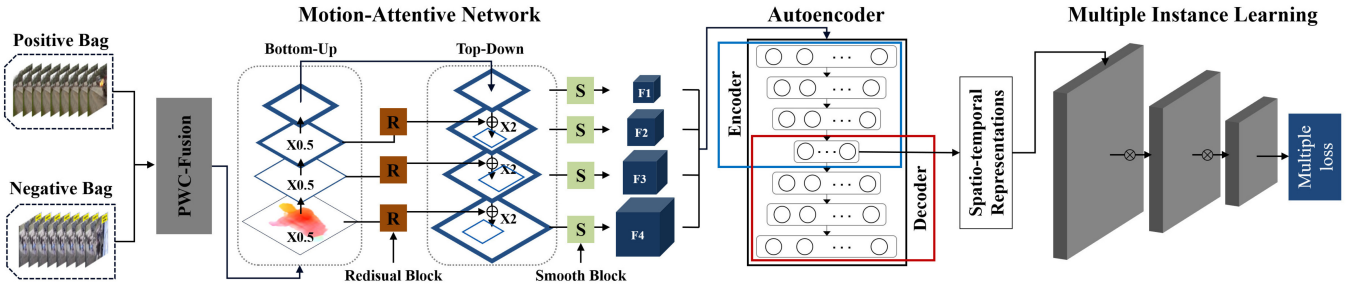

Figure 1: The flow diagram of the proposed motion-attentive network

ABSTRACT

Recently, numerous studies have utilized deep-learning-based approaches to detect anomalies in surveillance cameras. However, while several of these studies used motion features to detect abnormal situations, detection problems can arise due to the sparse information and irregular patterns in certain abnormal situations. We propose a means of preserving motion patterns in abnormal situations through a network called MA-Net, which solves representation problems caused by a loss of sparse information and irregular patterns. We show through experiments that the proposed method is superior to state-of-the-art methods.

CCS CONCEPTS

• **Social and professional topics** → **Surveillance**.

KEYWORDS

surveillance video, motion-attentive network, anomaly detection

ACM Reference Format:
U-Ju Gim, Jeong-Hun Kim, Kwan-Hee Yoo, and Aziz Nasridinov. 2020. Motion-Attentive Network for Detecting Abnormal Situations in Surveillance Video. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '20 Posters), August 17, 2020.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3388770.3407442

*Corresponding Author

SIGGRAPH '20 Posters, August 17, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7973-1/20/08.
https://doi.org/10.1145/3388770.3407442

1 INTRODUCTION

Surveillance cameras are used to detect abnormal situations though such detections are labor-intensive tasks. On the other hand, we can also consider such tasks as an anomaly detection problem, referring to the process of automatically identifying unexpected patterns from surveillance video. Existing approaches to detect abnormal situations use hand-crafted features [Hasan et al. 2016; Lu et al. 2013] or deep-learned features [Sultani et al. 2018; Zhu and Newsam 2019]. Hand-crafted feature-based approaches [Hasan et al. 2016; Lu et al. 2013] statistically extract appearance and motion patterns representing normal and abnormal situations from video frames and then classify abnormal situations through a deep-learning network. However, with hand-crafted features, it is difficult to detect abnormal situations, which typically present irregular patterns, even if they are of the same type. Deep-learned feature-based approaches [Sultani et al. 2018; Zhu and Newsam 2019] robustly classify abnormal situations by extracting spatiotemporal representations from irregular motion patterns using a deep learning network. Although the anomaly detection performance has dramatically improved with deep-learned feature-based approaches, it still requires further improvement. In particular, during the extraction of deep-learned features, relatively sparse information, which can be meaningful information to determine an abnormal situation, is lost. That is, a typical motion pattern for an abnormal situation may not always represent an abnormal situation, which leads to an abnormal detection failure.

To address this issue, we propose a motion-attentive network called MA-Net (see Figure 1). MA-Net learns enhanced deep-learned features that preserve significant portion of motion information related to abnormal situations in video frames. These enhanced deep-learned features solve the problem of failing to detect anomalies due to the loss of sparse information. We show through experiments that MA-Net outperforms state-of-the-art methods.

2 PROPOSED METHOD

First, we generate positive (abnormal) and negative (normal) bags containing video frames, and there is only video level label (see Figure 1). Next, we extract the optical flow using PWC-Fusion [Ren et al. 2019], which consists of motion representations for the video frames included in each bag. Subsequently, the optical flow for each bag is used as the input of the motion-attentive network.

The motion-attentive network extracts enhanced deep-learned features without a loss of meaningful information by preserving the attentive motion information, which helps to detect abnormal situations. To do this, we use feature pyramid network (FPN) [Lin et al. 2017], which combines feature maps of multiple scales extracted through convolution, as the backbone of this network. We configure the bottom-up pathway and top-down pathway of FPN with eight convolutional layers. Hence, it extracts a feature map for each convolution. We extract the feature maps of the compressed optical flow for each scale in the bottom-up pathway. Here, the most representative motion information of the video frame appears in the lower layer, and feature maps extracted for each scale are temporarily stored in the residual block and merged in the top-down pathway. In the top-down pathway, the network sequentially combines feature maps extracted from the bottom-up pathway. At this stage, the sparse information lost through convolution in the bottom-up pathway is restored. Thus, the final extracted features preserve the attentive motion information, which are our enhanced deep-learned features.

The autoencoder learns spatiotemporal representations from the enhanced deep-learned features extracted from motion-attentive networks. Here, spatiotemporal representations characterize patterns in which changes of the attentive motion information occur. Here, the input of the autoencoder is the enhanced deep-learned features corresponding to 16 continuous frames. Lastly, we adapt multiple-instance learning (MIL) [Sultani et al. 2018], which trains the autoencoder to undertake binary classification of abnormal situations in a weakly supervised manner. This allows the autoencoder to detect abnormal situations on a frame-by-frame basis.

3 EXPERIMENTS RESULTS

We use a large-scale real-world anomaly detection benchmark of the UCF Crime dataset [Sultani et al. 2018] to evaluate the proposed MA-Net. This dataset consists of a total of 1900 actual surveillance videos, with 950 containing 13 abnormal classes, including accidents, fights, shoplifting, thefts, and explosions. We use 1600 surveillance videos as the training dataset and the rest as the testing dataset. We measure the anomaly detection accuracy with the AUC (area under the curve) and compare MA-Net with state-of-the-arts approaches [Hasan et al. 2016; Lu et al. 2013; Sultani et al. 2018; Zhu and Newsam 2019]. Figure 2 is an example of visualizing the results of detecting two abnormal situations: abuse and car accident. From the Figure 2, we can observe that when using state-of-the-art [Sultani et al. 2018], anomaly detection fails when the sparse information representing an abnormal situation is lost. In contrast, MA-Net successfully detects abnormal situations by preserving attentive motion information and improving the AUC by more than 4% compared to deep-learned feature-based approaches [Sultani et al. 2018; Zhu and Newsam 2019], as shown in Table 1.

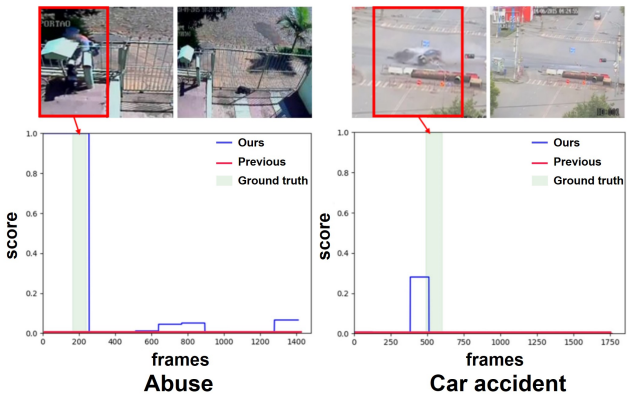

Figure 2: Visual examples of prediction results

Table 1: Performance comparison on the UCF Crime dataset

Model	AUC(%)
[Hasan et al. 2016]	50.6
[Lu et al. 2013]	65.5
[Sultani et al. 2018]	75.4
[Sultani et al. 2018] + [Zhu and Newsam 2019]	79.0
MA-Net (Ours)	83.1

4 CONCLUSION

In this paper, we proposed the MA-Net for detecting abnormal situations in surveillance video. The proposed MA-Net can detect abnormal situations based on enhanced deep-learned features by preserving sparse information. We evaluated the performance of MA-Net using the UCF Crime dataset, finding that the anomaly detection accuracy was significantly improved compared to existing state-of-the-art methods. This result is expected to be widely applied in the field of automated surveillance.

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

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*. 2720–2727.

Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik Sudderth, and Jan Kautz. 2019. A fusion approach for multi-frame optical flow estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2077–2086.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6479–6488.

Yi Zhu and Shawn Newsam. 2019. Motion-Aware Feature for Improved Video Anomaly Detection. *arXiv preprint arXiv:1907.10211* (2019).