Abnormal Human Action Detection Based on GAN

doi: https://doi.org/10.5954/ICAROB.2021.OS14-3

Abnormal Human Action Detection Based on GAN

Tomoya Sano, Seiji Ishikawa, Joo Kooi Tan

Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Sensui-cho, Tobata-ku, Kitakyushu 804-8550, Japan *

E-mail:{sano-tomoya,ishikawa, etheltan}@ss10.cntl.kyutech.ac.jp

Abstract

One of the important roles of a camera surveillance system is to detect abnormal human actions or events. In this study, we propose a method of abnormal human actions/events detection method using Generative Adversarial Nets (GAN). In anomaly action detection, the main problem is that the image data of abnormal human actions is more difficult to obtain than normal human actions. To solve this difficulty, we use only normal human action data in the employed training network and those actions not recognized as normal are judged as abnormal. Experimental results show effectiveness of the proposed method.

Keywords: anomaly detection, camera surveillance system, CNN, GAN, optical flow.

1. Introduction

Recently, surveillance systems using cameras have been widely used according to frequent outbreak of crimes. The number of surveillance cameras installed is on the rise. It is used in various places such as convenience stores, supermarkets, banks, stations, roads, and homes, and not only law enforcement agencies but also private companies and individuals exist as users. However, due to the shortage of manpower, it has not been possible to secure human resources to monitor surveillance camera images. Generally, the frequency of abnormal events is much lower than the frequency of normal events. As a result, most of the time and money is wasted on the work done by the observer. Therefore, there is an urgent need to develop a technology that automatically detects abnormal events from the images of surveillance cameras.

In this paper, we propose a detection method of an anomaly action on a video image using a convolutional neural network. As mentioned above, abnormal events rarely occur compared to normal events. Therefore, it is very difficult to collect data on abnormal events. We solve this problem by using only the data on normal events when training a model. Then an action not recognized as a normal action is judged as an abnormal action.

2. Acquisition of motion information

We estimate optical flows to capture motion features on a video. We use FlowNet [1] for this purpose. FlowNet is a deep convolutional neural network that provides a high-density optical flow for each estimated pixel, given two consecutive video frames. The estimated optical flow is provided in the form of a 2-channel (horizontal and vertical) map. Optical flow expresses the magnitude and direction of movement by a vector.

In the proposed method, we use the 2-channel map that takes the absolute value of the optical flow as motion

information. This prevents the loss calculated by the loss function during training (discussed later in Section 3.3) from becoming large due to the inversion of the positive and negative signs and promotes early convergence of the parameters. We call this feature expression Absolute Optical Flow (AOF) for convenience.

3. Framework

3.1. *GAN*

Generative Adversarial Nets (GAN) [2] is a framework for training generative models. It aims at finding a generative data distribution $p_g(\boldsymbol{x})$ that matches the true data distribution $p_{data}(\boldsymbol{x})$ obtained from the training data $\boldsymbol{x}$. Both the generator G, that captures the data distribution, and the discriminator D, that estimates the probability that the input data is obtained from $p_{data}(\boldsymbol{x})$, are trained while competing with each other.

The generator G generates the generated data $G(\boldsymbol{z})$ when random noise $\boldsymbol{z}$ according to the distribution $p_{\boldsymbol{z}}(\boldsymbol{z})$ is input to it. The input data $\boldsymbol{u}$ to the discriminator D is the training data $\boldsymbol{x}$ or the generated data $G(\boldsymbol{z})$. The output $D(\boldsymbol{u})$ of the discriminator D is the probability that the input data $\boldsymbol{u}$ belongs to the distribution of training data $p_{data}(\boldsymbol{x})$. In GAN, G and D are optimized by the value function $V(G, D)$ of the following equation;

$$\min_G \max_D V(G,D) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[\log\left(1 - D\left(G(\boldsymbol{z})\right)\right)\right] \quad (1)$$

In the above equation, G minimizes the objective function to trick D into generating data that is indistinguishable from true data. On the other hand, D finds the distinguishing boundary between the true data and the generated data by maximizing the objective function and avoids being fooled by G. In this way, G and D have a hostile relationship and are optimized while competing in GAN.

3.2. *Proposed framework*

Our framework learns only normal event data in surveillance camera images. We train two deep convolutional neural networks based on GAN. The first is generator G, which captures the data distribution of normal events. G gives one raw frame obtained from the camera and predicts the estimated motion information (AOF). The second is discriminator D, which estimates the probability that a given sample is AOF obtained by FlowNet. D discriminates whether the given input is AOF obtained by FlowNet or AOF estimated from one frame by G. Figure 1 shows the structure of the proposed framework.

Through this training, we aim at enabling the G to predict the appropriate motion information of the training data (normal event). We use the trained G during test. When the data of a normal action is given to G, the motion information prediction is performed appropriately. On the contrary, when the data including an abnormal action is given to G, the motion information is not estimated properly. As a result, the prediction result generated by G differs from the result obtained by FlowNet. In this way, anomaly detection is executed.

3.3. *Loss function*

The two models G and D minimize the following losses L_G and L_D by learning.

$$L_G = \text{softplus}\left(-D\left(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, G(\boldsymbol{x}_t)\right)\right) \quad (2)$$

$$L_D = \text{softplus}\left(-D\left(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, |F(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)|\right)\right) + \text{softplus}\left(D\left(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, G(\boldsymbol{x}_t)\right)\right) \quad (3)$$

$$\text{softplus}(n) = \log(1 - \exp n) \quad (4)$$

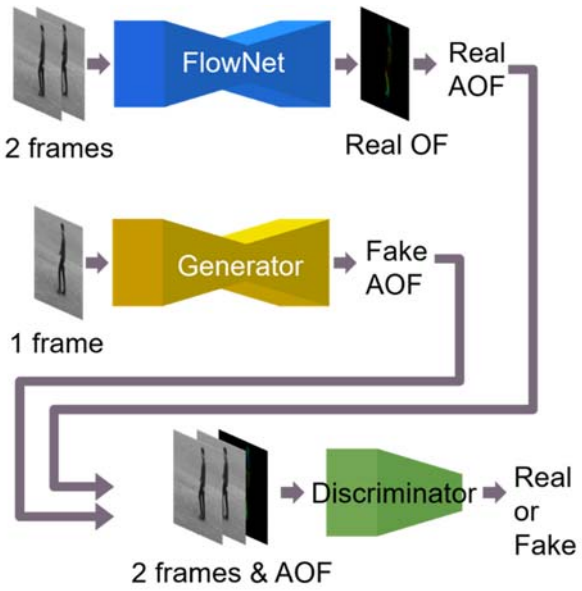

Fig. 1. The structure of the proposed framework.

where $G(\boldsymbol{x}_t)$ is the output of the generator G, $F(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$ is the output of FlowNet, $D\left(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, G(\boldsymbol{x}_t)\right)$ is the output of discriminator D that estimates the probability that the data generated by G belongs to the distribution of the estimation results by FlowNet.

3.4. *Anomaly detection*

Anomaly detection is executed every frame except the first frame. To detect anomalies, the squared error of the output of generator G and AOF obtained from FlowNet is calculated. The abnormal score at time t is calculated by the following equation.

$$E_t = \sum \left(G(\boldsymbol{x}_t) - F(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)\right)^2 \qquad (5)$$

When E_t exceeds a threshold value, the video frame at time t is detected as a frame containing an abnormal event.

3.5. *Network architecture*

3.5.1. *Generator G*

Generator G is a Deep Encoder-Decoder that has a skip coupling between the intermediate layer of the contracting part and the expansion part like U-net [3]. The input is one raw video frame (128×128×3). The output is AOF (128×128×2). Details are shown in Fig. 2.

In the 3×3 convolution layer, batch normalization is performed, followed by a rectified linear unit (ReLU). In the contracting part (upper side of Fig. 2), down-sampling is performed by 2×2 max-pooling, which doubles the number of channels in the feature map. After that, in the expansion part, on the contrary, up-sampling is performed by 2×2 deconvolution that halves the number of channels in the feature map.

3.5.2. *Discriminator D*

Discriminator D is a deep convolutional neural network with 10 convolutional layers. The input is a feature map (128×128×8) that connects two consecutive video frames and AOF. The output is the probability (scalar) that the AOF contained in the input is obtained by FlowNet. Details are shown in Fig. 3.

Discriminator D has the same structure as the encoder part of the generator G. Finally, it is output through a fully connected layer. It should be noted that no activation function is applied to the output layer.

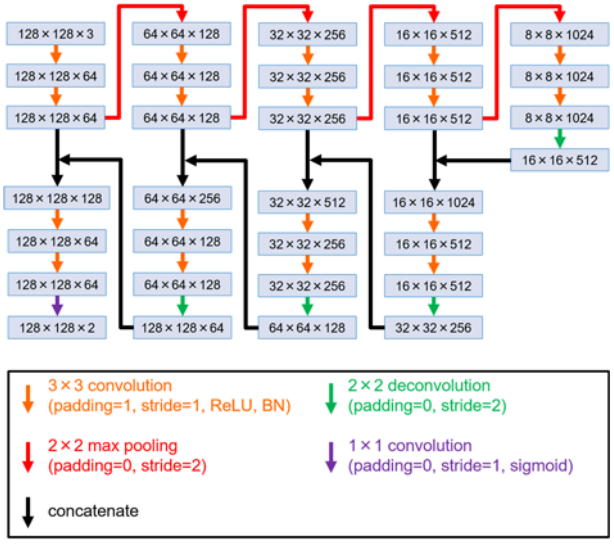

Fig. 2. Details of generator G. The blue block shows the feature map. The orange arrow indicates 3×3 convolution (padding=1, stride=1) with batch normalization, followed by a ReLU. The red arrow indicates 2×2 max-pooling with stride 2. The green arrow indicates 2×2 deconvolution with stride 2. The purple arrow indicates a 1×1 convolution with the sigmoid function applied. The black branched arrow indicates that the two feature maps are connected.

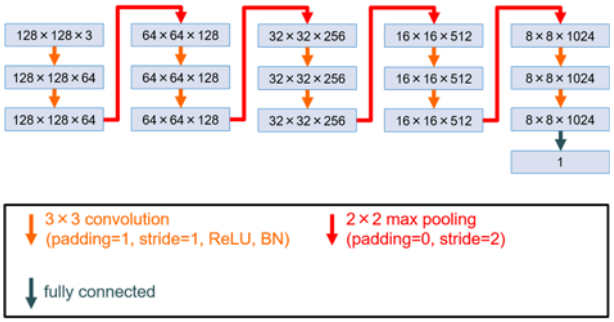

Fig. 3. Details of discriminator D. The dark green arrow indicates a fully connected layer. For other arrows, see the explanation in Fig. 2.

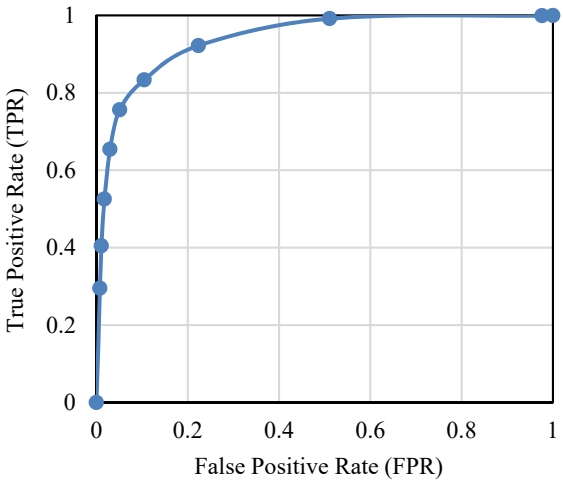

Fig. 4. ROC curve.

4. Experiment

Using the proposed framework, we perform a single-class motion classification experiment. For evaluation, a receiver operating characteristic (ROC) curve and an area under the curve (AUC) are employed. The ROC curve is obtained by calculating the true positive rate (TPR) and the false positive rate (FPR) for each of the normal and abnormal cutoff values for abnormality detection, and plotted on the coordinate plane with TPR on the vertical axis and FPR on the horizontal axis. The area under the curve is calculated to evaluate the effectiveness of the proposed method.

4.1. *Dataset*

The dataset used for the experiment is KTH dataset [4, 5]. KTH dataset is a human motion recognition dataset by 25 actors, including six motion classes: walking, jogging, running, boxing, hand waving, and hand clapping. This dataset contains videos shot under four different conditions (d1, d2, d3, d4), but in this experiment only the data under condition d1 is used. When inputting an image to the network, both ends are cut so that the image becomes square, and it is resized to a 128×128 pixel image. In walking, jogging, and running, there are frames with backgrounds that do not include people. In this experiment, such a frame is excluded in advance.

The walking class in this data set is treated as a normal class, and the other five classes of jogging, running, boxing, hand waving, and hand clapping are tested as abnormal classes. Out of the data of 25 actors, walking videos for 20 actors are used for model training. Six motion videos of walking, jogging, running, boxing, hand waving, and hand clapping by the remaining 5 people are used for the test.

4.2. *Result*

Figure 4 shows the ROC curve of a single-class motion classification experiment using a KTH dataset. The AUC calculated from Fig. 4 was 0.94.

5. Conclusion

We proposed a GAN-based method for detecting abnormal human action using only normal event data for training. Our experiments show that the proposed method is effective in detecting abnormal human actions. Future work includes application to more realistic surveillance video images and improvement of accuracy.

References

1. A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazırbas, V. Golkov, FlowNet: Learning Optical Flow with Convolutional Networks, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2758–2766.
2. I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, *Advances in neural information processing systems*, 2014, pp. 2672–2680.
3. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI*, 2015, pp. 234–241.
4. C. Schuldt, I. Laptev, Barbara Caputo, Recognizing Human Actions: A Local SVM Approach, *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004, Vol. 3, pp. 32–36.
5. KTH dataset: https://www.csc.kth.se/cvap/actions/
6. M. Ravanbakhsh, E. Sangineto, M. Nabi, N. Sebe, Training Adversarial Discriminators for Cross-Channel Abnormal Event Detection in Crowds, *Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1896–1904.
7. P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.