Abnormal Behavior Detection in Crowd Scene Using YOLO and Conv-AE

Li Yajing[1], Dai Zhongjian[2*]

1. Beijing Institute of Technology, Beijing 100080, China
E-mail: 458338022@qq.com

2. Beijing Institute of Technology, Beijing 100080, China
E-mail: padzj@sina.com
* Corresponding author: Dai Zhongjian (E-mail: padzj@sina.com)

Abstract: This paper proposes a weighted convolutional autoencoder (Conv-AE) and a novel regularity score based on the results of You Only Look Once (YOLO) network to detect abnormal behavior in crowd scenarios. The weighted Conv-AE extracts spatial features of video frames. In the training process, a weighted loss function is proposed based on the YOLO detection results, which emphasizes the foreground part, and thus overcomes the impact of complex background. In addition, a novel regularity score is put forward in the anomaly detection process. The regularity score takes into account the three factors of reconstruction errors obtained from weighted Conv-AE, speed information and category of objects detected by YOLO. Three scores respectively based on these factors are integrated to obtain anomaly detection results. The experimental results on UCSD ped1 and ped2 dataset verify that the proposed method achieves better performance than the most of semi-supervised methods.

Key Words: abnormal behavior detection, You Only Look Once (YOLO), convolutional autoencoder (Conv-AE), weighted loss function

2021 33rd Chinese Control and Decision Conference (CCDC) | 978-1-6654-4089-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCDC52312.2021.9602095

1 INTRODUCTION

Surveillance cameras are widely used in public areas to promote public security, resulting in large volumes of surveillance video data. In order to facilitate the understanding of these videos, many related research fields of intelligent monitoring system have arisen, such as object tracking [1], gait recognition [2], activity recognition [3] and abnormal detection [4], etc. Among them, abnormal detection has attracted the attention of more and more scholars due to its high relevance with public security, and has become the hottest issue in the field of video processing. The methods of abnormal behavior detection can be roughly divided into two categories: hand-crafted feature -based methods and deep learning-based methods [5].

Hand-crafted features-based methods extract low-level features to represent the video frame. Then the descriptors are used to build an abnormal behavior detection model. Commonly used descriptors include optical flow (OF) [6], spatial-temporal volume [7], trajectory [8], histograms of oriented gradients (HOG) [9], histograms of optical flow (HOF) [10], etc. Guo et al. [11] proposed a method combining mean shift and K-means classification to cluster optical flow from two aspects, so as to complete rapid and accurate crowd anomaly detection. Li et al. [12] proposed a joint detector of temporal and spatial anomalies based on a video representation that accounts for both appearance and dynamics using a set of mixture of dynamic textures models. Research on hand-crafted features-based methods appeared earlier, and there are many related studies. But in general, these methods have a common problem that human designed features inevitably bring noise and uncertainty, resulting in a relatively unsatisfactory effect.

With the great success of neural networks in computer vision in recent years, more and more researchers began to study abnormal behavior detection based on deep learning, and have achieved good results. Singh et al. [13] proposed Aggregation of Ensembles (AOE) structure of pre-trained ConvNets, where different CNN learn different levels of semantic representation from crowd videos. Sabokrou et al. [14] used fully convolutional neural networks (FCNs)and temporal data, transferred a pre-trained supervised FCN into an unsupervised FCN to detect and locate anomalies in videos. Generally, methods based on deep learning can get relatively better results than that based on hand-crafted features owing to the deeper features extraction.

However, the research on anomaly detection still has problems in terms of practical application. Firstly, there is no precise definition of abnormal behavior. Secondly, the abnormal behavior in videos is extremely incidental, that is, there is a serious imbalance between the positive and negative samples required to establish the detection model. In order to solve these issues, researchers recently proposed to train the anomaly detection models only based on normal behavior samples, which refers to semi-supervised learning models. AE [15] is widely used in anomaly detection because of its simple structure and training process. Yet the basic AE method may not achieve a great result due to the complex background and distraction of network.

This paper aims to solve this problem and obtain better detection results. Conv-AE is chosen as the basic network to extract the spatial characteristics of video frames considering the achievements of convolution in image processing. When coming to network distraction problem, a weighted loss function based on YOLO detection results is proposed to reduce the influence of background. Besides, motion features and object categories are also considered to obtain a novel regularity score.

The main contributions of our work are as follows:

1) A semi-supervised anomaly detection algorithm based on YOLO network and Conv-AE is proposed.

2) A weighted loss function based on YOLO is put forward to make the network pay more attention to the moving foreground, and thus reduce the influence of the complex background on the neural network.

3) The speed information and category of objects detected by YOLO are also taken into account, which lead to a more comprehensive scoring rule. Better detection results are obtained by integrating these two clues with the reconstruction error of Conv-AE.

The rest of this paper is organized as follows. Section 2 gives brief description of YOLO and Conv-AE as well as the related work. The proposed method is introduced in detail in Section 3. Experiments and results of this paper are included in Section 4. Finally, the conclusion of this paper and the direction for future studies is drawn in Section 5.

2 RELATED WORK

2.1. YOLO

YOLO [16] is a state-of-the-art, real-time object detection system proposed by Joseph Redmon et al. The prominent advantage of this method is the rapidity and high accuracy of target detection and recognition. The core idea of YOLO is to use the entire picture as the input of the network, and directly obtain the results of the bounding boxes, categories and confidence scores in the output layer.

In most cases, the YOLO network is mainly used as a part of object detection and tracking. However, thanks to its accuracy and rapidity, there are also some related studies that use YOLO to complete abnormal behavior detection. For example, YOLO network was used as the human detection module in Gong's [17] work. They obtained segmented patches of specific human subjects and delivered them into the 3D-CNN network to focus the network on learning motion characteristics of each person. Liu et al. [18] input the marked abnormal behavior directly into the YOLO network model for training without human target extraction, so as to realize the end-to-end abnormal behavior classification.

In this paper, YOLO v3 is selected to complete the target detection for the sake of following characteristics: 1) The use of multi-scale features for object detection enhances the ability to recognize small objects. 2) YOLO v3 improves the prediction accuracy while maintaining the speed advantage. In particular, the improve of small object recognition ability is very friendly to target recognition in crowd scenes. The results of bounding boxes, categories and confidence scores produced by YOLO network are applied in the proposed method.

2.2. Conv-AE

AE is a widely used deep learning method, which can automatically learn the hidden features from unlabeled data to complete pixel level tasks. AE is composed of an encoder and a decoder. In the task of abnormal behavior detection , the encoder encodes the input data to obtain a new implicit expression, which is decoded by the decoder to make the reconstructed output as similar as possible to the input. In the test process, the test samples that fail to fit the AE model are considered abnormal.

This paper chooses Conv-AE [19] network as the basic network, which uses convolutional layers and pooling layers instead of the fully connected layers as unit layers of encoder and decoder. The use of convolutional layers extends the application of AE from one-dimensional signals to two-dimensional images, which also makes Conv-AE widely used in the field of computer vision such as image noise reduction, dimension reduction feature extraction, and anomaly detection.

Hasan et al. [20] adopted a Conv-AE to perform anomaly detection. They completed two branch experiments. One was based on traditional manual extraction of features, the other was directly learning with the encoder. They compared and analyzed the results of two methods, and proved the feasibility of Conv-AE method in detecting abnormal behavior. In the work of Medel et al. [21], the proposed architecture used an AE structure comprised of Conv-LSTM (convolutional Long Short-Term Memory) units, which allowed the model to better learn the normality of a video. A new structure of Conv-AE was proposed by Luo et al. [22], applying Conv-LSTM layer after the encoder for temporal feature extraction. The current frame and the previous frame are both decoded in this method, which can more effectively use time information.

3 PROPOSED METHOD

3.1 Structure of proposed network

The structure of proposed method is shown in Fig 1. YOLO v3 is used in the data preparation process to accomplish the object extraction. The pre-trained YOLO network on the official website is chosen to complete the preprocessing part. The network was trained on the COCO dataset [23] and can detect 80 objects including people, bicycles, cars and so on, resulting in position of bounding box, category and related confidence score of each detected object. The results of YOLO v3 are utilized in both training and test process.

In the training process, the position and confidence score of each detected object are used to calculate the weight matrix and form the weighted loss function to train the Conv-AE. The calculation and usage of the weight matrix will be introduced in detail in Section 3.2. The reconstruction error of test frames trough Conv-AE is utilized as part of the proposed regularity score. Given the dynamic features of video, speeds of targets are calculated as an auxiliary information for calculating the regularity score. Motion feature extraction methods will be described in sections 3.3.

The category result from YOLO v3 detection is also used as a branch of regularity score calculation. Three branches mentioned above are integrated together and proposes a novel fused regularity score to complete the abnormal detection. The calculation of the proposed regularity score will be mentioned in section 3.4.

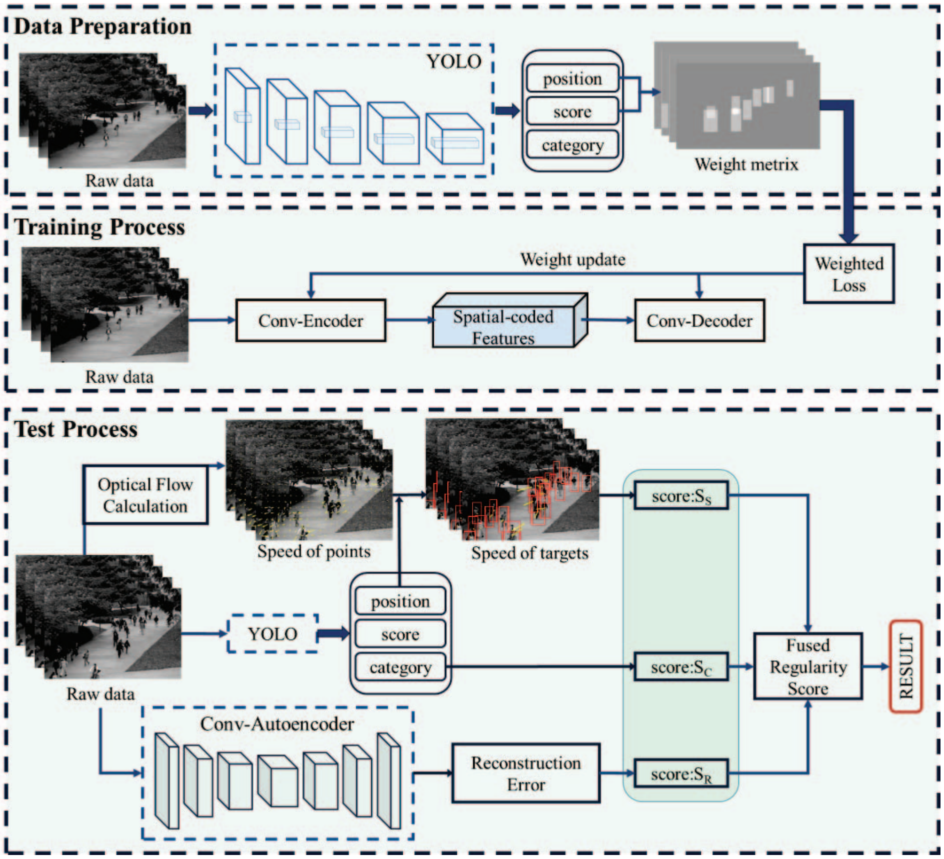


Fig 1. Structure of proposed algorithm.

Fig 2 shows the structure and the parameters of Conv-AE network. In the convolution layer and the deconvolution layer, the parameter is expressed as filter number @ kernel size - step - padding. The parameters of max pooling layer and up-sampling layer are pool size and up-sampling size, respectively.

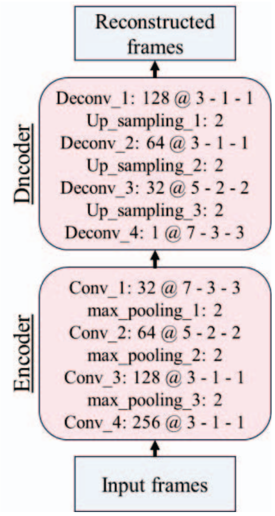


Fig 2. Structure and parameters of Conv-AE.

3.2 Weighted loss function based on YOLO

In order to weaken the distraction of the network, the detection results of the YOLO v3 is used to form a new network loss function. Since the background does not help much for the detection of abnormal behavior, the proposed weighted loss function weakens the background and emphasize the foreground, which is defined as:

$$L = \sum_{x,y} (p_{xy}^{rec} - p_{xy}^{inp})^2 w_{xy} \tag{1}$$

where p_{xy}^{rec} and p_{xy}^{inp} are the pixel values of the reconstructed image and the input image at point (x, y), and w_{xy} is the weight matrix of the point. The size of the weight matrix is consistent with the size of the input and output images. The calculation formula of w_{xy} is as follows:

$$w_{xy} = 1 + \sum_{i=1}^{n_k} score_i \ , (x, y) \in i \tag{2}$$

where n_k is the number of boxes detected in frame k, $score_i$ is the confidence score of the i-th detected object. If the pixel point (x, y) is included in the YOLO detection box i, the weight of the point is increased by $score_i$.

3.3 Extraction of motion features

The Conv-AE can only extract spatial features of each frame, but it is obvious that motion features are equally significant for abnormal behavior detection. This paper will extract motion features as well to detect abnormal behavior.

Optical flow is a method that can effectively extract motion features. In this paper, Lucas-Kanade optical flow [24] method is used to calculate optical flow field. Assuming that the gray value of point (x,y) in frame k is $I(x,y,k)$. After Δk frames, the point (x,y) moves to $(x+\Delta x, y+\Delta y)$, and the gray value is $I(x+\Delta x, y+\Delta y, k+\Delta k)$. The optical flow method assumes that after Δk frames, the gray value of point (x,y) changes little, and the optical flow constraint equation is shown in equation (3):

$$\frac{\Delta x}{\Delta k} I_x + \frac{\Delta y}{\Delta k} I_y + \frac{\Delta k}{\Delta k} I_k = 0 \tag{3}$$

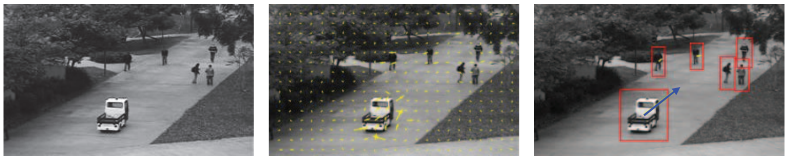
where I_x, I_y and I_k are the partial derivatives of the gray value function $I(x,y,k)$ to the variables x, y and k. The velocity in the x and y directions at the frame k, i.e., $V_x(x,y,k)$ and $V_x(x,y,k)$ can be solved through the constraint equation, and the optical flow diagram about the crowd motion can be obtained.

Considering that the speed of most abnormal behaviors in pedestrian scenarios is larger than normal, we further find the speed of each individual detected object as follows:

$$|V(i,k)| = \sqrt{(\sum V_x(x,y,k))^2 + (\sum V_y(x,y,k))^2}, \quad (x,y) \in i \tag{4}$$

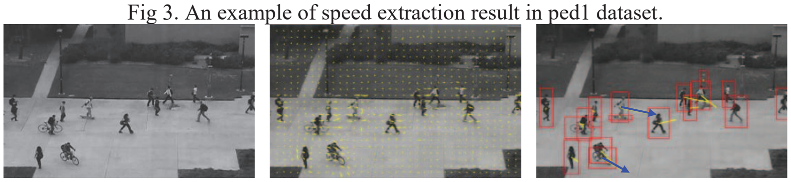
where $|V(i,k)|$ is the speed of detected object corresponding to the *i-th* detection box of YOLO in frame k, $i \in [1, n_k]$.

Fig 3 and Fig 4 show the speed extraction results of frames in ped1 datasets and ped2 datasets respectively. In each picture, Fig(a) represents the original image of the frame, Fig(b) and Fig(c) represent the motion information of equidistant sampling points and detected objects, respectively. The yellow arrows represent the direction and the amplitude of movement. In addition, we use the blue arrow in Fig(c) to emphasize the faster speed caused by abnormal behavior in the frame. It can be seen that the abnormal behavior caused by a car in Fig 3 and a bike and a skater in Fig 4 have higher speed.



(a) (b) (c)

Fig 3. An example of speed extraction result in ped1 dataset.



(a) (b) (c)

Fig 4. An example of speed extraction result in ped2 dataset.

3.4 Calculation of the fused regularity score

The regularity score fusion process is shown in the test process of Fig 1. The fused regularity score consists of three parts: the score obtained from the AE reconstruction error S_r, the score obtained from the extracted speed feature S_s, and the score obtained from the classification result detected by YOLO v3 S_c.

For the *k-th* frame, the reconstruction error is obtained by:

$$e(k) = \sum_{x,y} (p_{xy}^{rec} - p_{xy}^{inp})^2 \tag{5}$$

Then the reconstruction regularity score is calculated by normalizing the error:

$$S_r(k) = 1 - \frac{e(k) - \min_k e(k)}{\max_k e(k) - \min_k e(k)} \tag{6}$$

As for the motion feature, we use the maximum speed in all of detected objects in each frame as an indicator to calculate the speed regularity score.

$$S_{s\_clip}(k) = 1 - \frac{|V(k)| - \min_{k\_clip} |V(k)|}{\max_{k\_clip} |V(k)|} \tag{7}$$

where $|V(k)| = \max_i |V(i,k)|$ is the maximum speed in frame k, $\min_{k\_clip} |V(k)|$ and $\max_{k\_clip} |V(k)|$ correspond to the maximum and minimum values of $|V(k)|$ in a video clip.

Considering that some video clips are all marked as abnormal, we also calculated the speed score relative to the entire dataset:

$$S_{s\_all}(k) = 1 - \frac{|V(k)| - \min_{k\_all} |V(k)|}{\max_{k\_all} |V(k)|} \tag{8}$$

Where $\max_{k\_all} |V(x)|$ and $\min_{k\_all} |V(k)|$ is the maximum and minimum values of $|V(k)|$ in a dataset, respectively.

$S_{s\_clip}(k)$ reflects the relative value of the maximum speed in a frame to the video clip. While $S_{s\_all}(k)$ reflects the relative value of the maximum speed in a frame to the whole dataset. A small $S_{s\_all}(k)$ value indicates that the maximum speed of the frame is large relative to the whole dataset, so the speed regularity score of the frame whose $S_{s\_all}(k)$ value is less than the threshold value is set to zero, and determine it as an abnormal frame. Then we can get the speed regularity score as:

$$S_s(k)=\begin{cases}0 & ,S_{s\_all}(k)<0.8\\ S_{s\_clip}(k) & ,S_{s\_all}(k)\ge 0.8\end{cases} \qquad (9)$$

The categories obtained by the detection results is also applied to the calculation of class regularity score. Consider the sidewalk scene, if there are objects detected like bicycles, cars or skates in a frame, then the class regularity score $S_c(k)$ is set to zero, which can directly determine that the frame is abnormal. If only humans are detected in a frame, the class regularity score $S_c(k)$ is set to one, and the final score is determined by the other two scores. Thus, the class regularity score is determined as Eq. 10, which is as a multiplier.

$$S_c(k)=\begin{cases}0 & \textit{if} \text{ other categories in result}\\ 1 & \textit{if} \text{ only people in result}\end{cases} \qquad (10)$$

Finally, the regularity score of the frame k is defined. As mentioned above, $S_c(k)$ is the multiplier. As an auxiliary condition, the speed regularity score $S_s(k)$ is mainly sensitive to the abnormal behavior with large speed, that is, the possibility of abnormal behavior with small $S_s(k)$ value is relatively large. Based on the above analysis, the final results are defined as follows:

$$S(k)=\begin{cases}S_c(k)*\min(S_r(k),S_s(k)) & ,S_s(k)<0.8\\ S_c(k)*S_r(k) & ,S_s(k)\ge 0.8\end{cases} \qquad (11)$$

The regularity score $S(k)$ is used to determine the abnormal frames. The lower the score is, the more likely the frame is to have abnormal behavior.

4 EXPERIMENTS

The proposed algorithm is implemented using TensorFlow and Keras framework with Python on Ubuntu operating system, and experiments are conducted on a server equipped with NVIDIA GTX2080 SUPER GPU.

4.1 Data Description

To validate our proposed method, we tested our algorithm on UCSD dataset [25], which is commonly used in branches of abnormal behavior recognition, and consists of two sub-data sets called ped1 and ped2. Both of these two datasets contain video frames in pavement scenarios recorded with a stationary camera. The anomalies in the dataset are defined as cars, bickers and skaters.

Ped1 contains a training set with 34 normal video sequences and a test set with 36 abnormal video sequences. Each video sequence consists of 200 frames of which the size is 158*238. Ped2 contains a training set with 16 normal video sequences and a test set with 12 abnormal video sequences. Each video sequence consists of 120, 150 or 180 frames of which the size is 240*360.

4.2 Detection results

The proposed method is completed to detect anomaly on UCSD dataset and the results are displayed in this section. Fig 5 and 6 are the detection results of proposed algorithm on a video segment in the ped1 and ped2, respectively.

As shown in figure, the blue curve represents the regularity score obtained by our method and the red one represents the ground truth. For each frame in video clips, if the frame contains abnormal behavior, the value of ground truth is 0, otherwise is 1. We choose normal and abnormal frames according to the result, and use yellow boxes indicating the abnormal behavior in abnormal frames, which are caused by skater and bicycle, respectively.

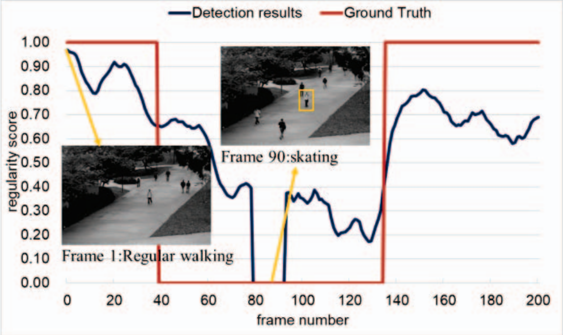


Fig 5. Video-level anomaly detection in the 25th video clip of ped1.

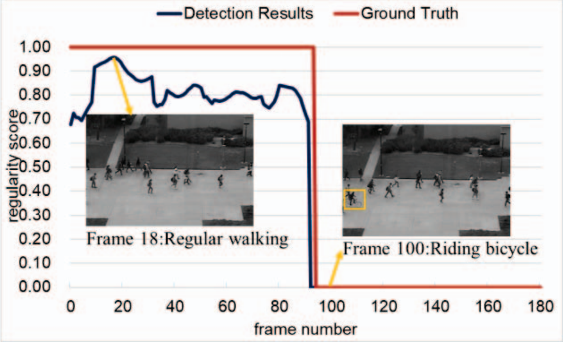


Fig 6. Video-level anomaly detection in the 2-nd video clip of ped2.

The results obtained from the proposed method is compared with some semi-supervised algorithms. Area under the curve (AUC) is a commonly used index to evaluate the results. In general, the higher the AUC value is, the better the effect of the algorithm is and the more accurate the detection is.

Table 1. Comparison with the state-of-the-art semi-supervised methods in terms of AUC% (area under curve) on UCSD Ped1 and Ped2 dataset

Method	Ped1	Ped2
Conv-AE [20]	68.1%	81.1%
Conv-LSTM-AE [22]	75.5%	88.1%
MGFC-AAE [26]	**85.0%**	91.6%
Proposed method	83.8%	**92.8%**

We compare the proposed method with semi-supervised and deep learning-based approaches. The results are shown in Table 1. It can be seen from the table that our method gets good results when compared with the three other methods based on deep learning. Our method improves 13.7% and 11.7% for AUC on ped1 and ped2 dataset respectively when compared to the basic Conv-AE [20]. It also improves 6.3% and 4.7% for two datasets for Conv-LSTM-AE [22], which integrated Conv-AE and LSTM with Auto-Encoder. The MGFC-AAE [26] used the original image and optical flow image as input data to build a dual stream AE. This method achieved a little better result than ours on ped1, but we get an increase of 1.2% on ped2, which can prove the validity of our method.

We also conduct comparative experiments with the same network and parameters to prove the effectiveness of the proposed speed regularity score and class regularity score. Results are shown in Table 2.

Table 2. Validation of different regularity score in terms of AUC% (area under curve) on UCSD Ped1 and Ped2 dataset

Type of Regularity Score	Ped1	Ped2
Reconstruction	70.7%	82.0%
Reconstruction + Speed	73.6%	84.2%
Reconstruction + Class	77.1%	91.8%
Reconstruction + Speed + Class	**83.8%**	**92.8%**

5 CONCLUSION

This paper proposes a novel semi-supervised method for abnormal behavior detection based on Conv-AE. In order to weaken the impact of the background, the weighted loss function based on YOLO detection is applied to the training process of Conv-AE. In addition, the speed characteristics and categories of detected objects are used as auxiliary clues and a new regularity score is put forward to complete abnormal behavior detection. The experiment results on two datasets demonstrate the effectiveness of our method, and the comparison with other semi-supervised methods can also reflect the superiority of this method. However, since categories are used as auxiliary information to determine anomalies, this method is more suitable for pedestrian scenarios. At the same time, there is room for further improvement in the accuracy of the method. In future work, we will continue improving the algorithm to make the generalization ability and algorithm performance better.

REFERENCES

[1] T. Yang, C. Cappelle, Y. Ruichek, and M. E. Bagdouri. Online multi-object tracking combining optical flow and compressive tracking in Markov decision process. Journal of Visual Communication & Image Representation, 58:178-186, 2019.

[2] J. Figueiredo, C. P. Santos, and J. C. Moreno. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. Medical Engineering & Physics, 53:1-12, 2018.

[3] M. Ma, N. Marturi, Y. Li, A. Leonardis, and R. Stolkin. Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognition, 76:506-521, 2018.

[4] A. Patcha, and J-M. Park . An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks, 51:3448-3470, 2007.

[5] A. B. Mabrouk, and E. Zagrouba. Abnormal behavior recognition for intelligent video surveillance systems : a review. Expert Systems with Applications, 91:480-491, 2018.

[6] X. Chen, and J. Lai. Detecting abnormal crowd behaviors based on the div-curl characteristics of flow fields. Pattern Recognition, 88:342-355, 2019.

[7] R. Chaker, Z. A. Aghbari, and I. N. Junejo. Social Network Model for Crowd Anomaly Detection and Localization. Pattern Recognition, 61:266-281, 2017.

[8] H. Fradi, B. Luvison, and Q. C. Pham. Crowd Behavior Analysis Using Local Mid-Level Visual Descriptors. IEEE Transactions on Circuits and Systems for Video Technology, 27(3):589-602, March 2017.

[9] N. Dalal, and B. Triggs, Histograms of oriented gradients for human detection, in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 886-893, 2005.

[10] N. Dalal, B. Triggs, and C. Schmid, Human detection using oriented histograms of flow and appearance, in Proceedings of the European Conference on Computer Vision (ECCV), 428–441, 2006.

[11] S. Guo, Q. Bai, S. Gao, Y. Zhang, and A. Li. An Analysis Method of Crowd Abnormal Behavior for Video Service Robot. IEEE Access, 7:169577-169585, 2019.

[12] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly Detection and Localization in Crowded Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(1):18-32, 2014.

[13] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia. Crowd anomaly detection using Aggregation of Ensembles of fine-tuned ConvNets. Neurocomputing, 371:188-198, 2020.

[14] M. Sabokrou, M. Fayyaz1, M. Fathya, Z. Moayedc, and R. Klettec. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, 172:88-97, 2018.

[15] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding, 195:1-12, 2020.

[16] Redmon, Joseph and Farhadi, Ali. YOLOv3: An incremental improvement. arXiv, 2018.

[17] M. Gong, H. Zeng, Y. Xie, H. Li, and Z. Tang. Local distinguishability aggrandizing network for human anomaly detection, Neural Networks, 122:364-373, 2020.

[18] X. Li , and S. Sun. Research on abnormal behavior detection based YOLO network. Electronic Design Engineering, 26(20):154-158,164, 2018.

[19] J. Masci, U. Meier, D. Ciresan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. in Proceedings of the international Conference on Artificial Neural Networks (ICANN), 52-59, 2011.

[20] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 733–742, 2016.

[21] J. R. Medel, and A. Savakis, Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv, 2016.

[22] W. Luo, W. Liu, and S. Gao, Remembering history with convolutional LSTM for anomaly detection. in Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 439–444, 2017.

[23] T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. in Proceedings of the European Conference on Computer Vision (ECCV), 740–755, 2014.

[24] B. K. P. Horn, and B. G. Schunck. Determining optical flow. Artificial Intelligence, 17(1-3):185-203, 1981.

[25] A. B. Chan, Z-S. J. Liang, N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1-7, 2008.

[26] N. Li, F. Chang, Video anomaly detection and localization via multivariate Gaussian fully convolution adversarial autoencoder. Neurocomputing, 369:92-105, 2019.