Multiscale feature fusion for surveillance video diagnosis

Article *in* Knowledge-Based Systems · January 2022

DOI: 10.1016/j.knosys.2021.108103

CITATION
1

READS
49

5 authors, including:

Fanglin Chen
Harbin Institute of Technology, Shenzhen
35 PUBLICATIONS **726** CITATIONS
SEE PROFILE

Wenjie Pei
Harbin Institute of Technology Shenzhen Graduate School
48 PUBLICATIONS **474** CITATIONS
SEE PROFILE

Knowledge-Based Systems 240 (2022) 108103

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Multiscale feature fusion for surveillance video diagnosis

Fanglin Chen, Weihang Wang, Huiyuan Yang, Wenjie Pei *, Guangming Lu

Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong, 518055, China

ARTICLE INFO

Article history:
Received 1 November 2021
Received in revised form 14 December 2021
Accepted 30 December 2021
Available online 7 January 2022

Keywords:
Surveillance video diagnosis
Anomaly classification
Multiscale feature fusion
Deep learning

ABSTRACT

Recently, surveillance video diagnosis has attracted increasing interest for generating real-time alarms related to camera failure in video surveillance systems. The existing surveillance video diagnosis methods do not have sufficient ability to detect multiple types of anomalies. Therefore, this paper proposes a surveillance video diagnosis method based on deep learning to detect multiple types of anomalies. A multiscale feature fusion residual network is designed to detect and classify camera anomalies. The experimental results show that the classification accuracy of the proposed method is more than 98%.

1. Introduction

With the growing demand for city security, the number of cameras installed in video surveillance networks is increasing, and the service time is gradually evolving to an uninterrupted running time of 24/7 [1,2]. Cameras are usually installed in exposed environments; thus, they suffer risks such as natural damage, man-made destruction and equipment failure. The video quality may be affected by blurred images, anomalous color casts, occlusion, and extremely bright or dark images. If the video suffers quality anomalies, the video content analysis methods for safety and security will be affected [3]. The important premise of behavior recognition and intelligent analysis is to ensure the image quality of surveillance videos. Thus, there is huge demand for detecting and classifying surveillance video anomalies, which is called surveillance video diagnosis.

Generally, there are two basic methods to diagnose video quality: subjective diagnosis and objective diagnosis. Subjective diagnosis is more effective because the human visual system is powerful. However, the expensive cost of human resources confines its application in diagnosing large-scale video surveillance systems. Moreover, manual monitoring of surveillance systems is tedious, and research shows that the efficiency of subjective diagnosis is very low [4]. To overcome these shortcomings, objective diagnosis has been studied in recent years.

There are two kinds of objective diagnosis technologies: image based and video based. The image based technique uses one frame of the camera to predict the anomaly, while the video based technology uses a short video to detect the anomaly. Usually, the video based methods are more robust than the image based ones. For example, motion analysis [5,6] can be used for still frame detection. However, video based methods are more complicated and time consuming, and they are not suited for large video surveillance networks which have millions of cameras. This paper focuses on image based diagnosis technologies, since they can classify anomaly by using one frame and are more effectively. What is more, a diagnosis system can use image based method more than one time (use several frames and vote) to improve the robustness. For example, a common phenomenon is that the camera occasionally has a focus error causing blur, but it will recover immediately. In this situation, the detection result of sequential sampled frames can vote that the camera is normal.

Image base objective diagnosis can be generally divided into two categories: traditional methods and deep learning-based methods. Traditional manually designed feature [7] methods need to diagnose different anomalies one by one. Usually, a set of thresholds is used to detect whether there are anomalies. Such rule-based methods can only work effectively in a specific scenario, and the accuracy of the technique will decline rapidly due to the change in scenes. Thus, the effect of the traditional method highly depends on the application scenario, and the generalization ability has difficulty achieving expectation.

To address more types of camera anomalies, a camera anomaly detection and classification method based on deep learning is proposed in this paper. In recent years, deep learning has developed rapidly in the field of computer vision and has achieved excellent results in the fields of face recognition [8], image classification [9], behavior recognition [10,11] and background restoration [12]. Deep learning methods can save the steps of manually designing features and automatically learn features from a large number of datasets, which is the largest difference between them and the traditional methods. Deep learning can simulate human neurons for data analysis and processing [13]. It can establish a

* Corresponding author.
E-mail address: wenjiecoder@outlook.com (W. Pei).

https://doi.org/10.1016/j.knosys.2021.108103
0950-7051/

very complex model to make the expression ability of features more effective (this can improve the generalization ability of features) and is able to deal with complex surveillance scenes. It usually has a special deep structure and can validly learn the complex mapping between input and output.

Therefore, surveillance video anomaly diagnosis based on deep learning can extract more effective features and improve the accuracy of classifying anomaly types. This paper aims to employ the deep learning method to detect and classify anomalies in surveillance videos to improve the applicability and robustness of surveillance video anomaly detection for different scenes. The contributions of this paper are as follows:

- We propose an image based surveillance video diagnosis technique for multiclass anomaly classification.
- A multiscale feature fusion residual network is proposed to extract both the global and local features of the image. The global feature can predict anomalies that have "unitary" changes, such as color_cast and brightness. The local feature is capable of distinguishing anomalies which have "intricate" variations, such as blur and occlusion.
- Experimental results show that the proposed algorithm can gain a classification accuracy of 98.52%, which indicates the effectiveness of the proposed technique.

This paper focus on detecting the anomalies of the camera itself, such as anomalies caused by blurred images, anomalous color casts, occlusion, and extremely bright or dark image. The detection of abnormal active events in videos are not studied in this paper. The rest of this paper is organized as follows. Section 2 gives an overview of related works in this area. Section 3 describes the proposed method in detail. The experimental results and discussion are given in Section 4. Finally, Section 5 draws the conclusions.

2. Related works

Surveillance video diagnosis has been widely studied. Almost all of these research methods are based on lightweight computer vision kernels [14], such as histogram-based analysis, feature thresholding, and background comparison.

Histogram-based analysis is widely used for surveillance video diagnosis, such as occlusion detection. Aksay et al. [15] calculated the absolute difference histogram of the frame and the background image and then compared the sum of the first k bins with the sum of all bins. Saglam et al. [16] proposed a histogram that calculated for the brightness level of the current frame and the background image, and the maximum histogram value was used as a feature. The works in [17,18] took the same idea but used the sum of the maximum neighbors in the histogram for occlusion detection.

Feature thresholding is the most common anomaly detection method. The threshold can be absolute or relative to the normal value of the features that can be calculated from the previous frame or a background image. The representative methods include the mean square error (MSE), peak signal noise ratio (PSNR) [19] and weighted mean square error. The above methods are usually used in snow point interference detection. Traditional blur detection methods of manually extracting features usually use the local gradient, frequency, singular value and other manual features of the image, which are concentrated in the gradient domain, frequency domain or other transformation domain.

Feature extraction methods based on gradients have been widely studied and applied. Levin et al. [20] distinguished the blur region and distinct region of an image by statistical gradient direction information and realized more accurate blur division for complex blur scenes with different blur kernels superimposed since the method was unable to be modeled with a single kernel function. Kim and Lee et al. [21] combined the gradient amplitude and direction consistency features, used a classifier (e.g., SVM, LDA [22]) to distinguish each pixel into clear, defocus blur or motion blur, and used the superpixel segmentation technology to divide the image into continuous clear, defocus blur and motion blur areas.

In terms of the feature extraction method in the frequency domain, Golestaneh et al. [23] proposed a blur detection method based on multiscale high-frequency fusion, which used the high-frequency discrete cosine transform coefficients extracted from multiple resolution image blocks for blur detection. Javaran et al. [24] used discrete cosine transform coefficients to describe the blur intensity of each pixel and distinguished the blurred region from the clear region based on pixel theory. In [16], the high-frequency components extracted by Fourier transform are the features used to detect camera defocus.

In addition, there are some blur detection studies based on other mechanisms. Yi et al. [25] proposed an image definition measure based on a local binary pattern (LBP) to distinguish the focused or defocused areas in the image. Pang et al. [26] proposed a specific blur kernel feature vector for blur detection by multiplying the filtered blur kernel variance and the filtered image block gradient variance. Shi et al. [27] distinguished clear pixels from blurred pixels by combining spatial domain filter, Fourier domain descriptor and local gradient edge distribution.

Background comparison was used in [28] for occlusion detection. It proposed using short-term and long-term background images. The short-term images were updated with the current frame, and the long-term images were updated with the short-term model. They computed the histograms using only the fixed pixels for long-term and short-term background images. Then, the correlation between these histograms was used for occlusion detection. In [29], the background subtraction method was used for occlusion detection. The background model was first established, and the foreground object was then interpreted as occlusion. Ribnick et al. [30] proposed identifying camera anomalies by detecting the large differences between new and previous video frames.

With the development of deep learning [31], a few works based on neural network have been proposed for surveillance video diagnosis. Dong et al. [1] proposed a neural network with four convolution layers and one full connection layer to detect occlusion and defocus events. Liang et al. [32] built a model based on neural network and optimization driven support vector machine for soft multimedia anomaly detection.

3. Methods

In this section, we propose to use deep learning approaches for surveillance video diagnosis. In our approach, we detect surveillance video anomalies using only one frame and convert the surveillance video diagnosis problem to a multiclass task for common surveillance video image anomalies such as brightness anomaly, blur, occlusion, etc. We propose a multiscale feature fusion residual network for this task. Though our approach is designed in an image-based manner, it can be readily used in a video sequence by sampling frames and feeding them into our system one by one. Then the final result is aggregated by voting from all the predictions of the sampled frames spanning a temporal interval. Thus our method is robust to occasional errors. On the other hand, the long-term anomaly can be reflected from the aggregated prediction. In the rest of this section, we first describe the feature extraction residual network and then explain the multiscale feature fusion mechanism.

Table 1
CNN architecture for surveillance video diagnosis.

Layer No.	Layer type	Layer size	Description
1	Input layer	RGB image	Input image is scaled to a size of 448 × 448
2	CONV	3 feature maps, each with a size of 224 × 224	Convolution using a 7 × 7 window and stride of 2
3	CONV	64 feature maps, each with a size of 112 × 112	Convolution using a 7 × 7 window and stride of 2, followed by batch normalization and ReLU activation
4	POOL	64 feature maps, each with a size of 56 × 56	Maxpooling with a 3 × 3 window and stride of 2
5	Residual block	64 feature maps, each with a size of 56 × 56	The first and second convolution in the block both use a 3 × 3 window and stride of 1
6	Residual block	128 feature maps, each with a size of 28 × 28	The first convolution in the block uses a 3 × 3 window and stride of 2, and the second uses a 3 × 3 window and stride of 1
7	Residual block	256 feature maps, each with a size of 14 × 14	The first convolution in the block uses a 3 × 3 window and stride of 2, and the second uses a 3 × 3 window and stride of 1
8	Residual block	512 feature maps, each with a size of 7 × 7	The first convolution in the block uses a 3 × 3 window and stride of 2, and the second uses a 3 × 3 window and stride of 1

3.1. Residual network for feature extraction

Convolutional neural network (CNN), an important representative in the field of deep learning, has been widely used in many fields in recent years, such as face recognition, speech conversion, image classification, and background restoration [33]. Most convolutional neural networks have network structures such as convolution layer, pooling layer and full connection layer. They have strong learning ability and can express and learn very complex features well. Since the proposal of AlexNet in 2012, a variety of convolutional neural network algorithms have gradually appeared, including VGG, Inception, ResNet, etc. Residual network has been continuously improved over time in theoretical research and has achieved good landing results in practical applications. In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), various architectures were evaluated on the large-scale ImageNet image database. There are more than 14 million images in the dataset and more than 21 thousand groups or classes (synsets). ResNet won the challenge in 2015.

In this study, we proposed a residual network for the task of surveillance video diagnosis. We used 2 convolutional (CONV) and 4 residual blocks. The complete architecture is described in Table 1. The input of the network is a four-dimensional tensor (N, C, H, and W), and N is the tensor quantity of each batch. C is the number of channels, and the surveillance video images used in this paper have 3 channels (RGB — Red, Green, Blue). H and W are the input dimensions, and all of the input images are scaled to 448 × 448.

The input dimension of most popular network is 224 × 224. If we use this in the data preprocessing and resize the 1920 × 1080 surveillance image to 224 × 224, the image reduction ratio will be too large. This makes it difficult to observe blurred anomaly in the original fuzzy and noisy surveillance video image, and the fuzzy and noise characteristics are lost. Therefore, we considered improving the network structure and increasing the input size of the network to better retain the fuzzy features and improve the surveillance image blur anomaly diagnosis ability of the model.

In the 2 convolutional layers, 3 and 64 feature maps are used, respectively. We use a 7 × 7 window and stride of 2 to perform the convolution operation. Batch normalization and ReLU activation function are applied after each convolution layer to avoid overfitting, obtain better optimization performance and speed up the training process [34,35]. The 2 convolutional layers change the feature size to 64 × 112 × 112. Max pooling layer is processed after the second convolutional layer, and the size of the feature maps is reduced from 64 × 112 × 112 to 64 × 56 × 56.

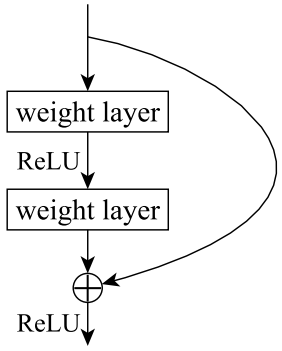

Fig. 1. The framework of the residual block.

In the 4 residual blocks, shortcut connections that simply perform identity mapping are applied [36], and their outputs are added to the outputs of the stacked *weight* layers (Fig. 1 shows the framework of the residual block). The functional relationship of the residual block is as follows:

$$F(x) = H(x) - x, \tag{1}$$

where $H(x)$ is the output of the residual block and $F(x)$ is the output after the intermediate stacked layers. When the problem of vanishing gradients occurs during backpropagation, the function can make the original input as output by identity mapping, which is equivalent to the process of input feature replication and transfer. Through this processing, the model training will not be adversely affected by vanishing gradients.

The stacked layers in all four residual blocks are all convolutional layers with a 3 × 3 window. In the first residual block, the stride of the two convolutional layers is 1, and the number of channels remains unchanged; thus, the output is still 64 × 56 × 56. However, in the 2nd–4th residual blocks, the first convolution in the block uses stride of 2, and the second still uses stride of 1. We also double the number of output channels in the first convolutional layer. To unify the number of channels in the layer during residual connection, a downsampling operation and channel number doubling are added to the shortcut connection. The outputs of 2nd–4th residual blocks are 128 × 28 × 28, 256 × 14 × 14 and 512 × 7 × 7, respectively, which extract deeper features gradually.

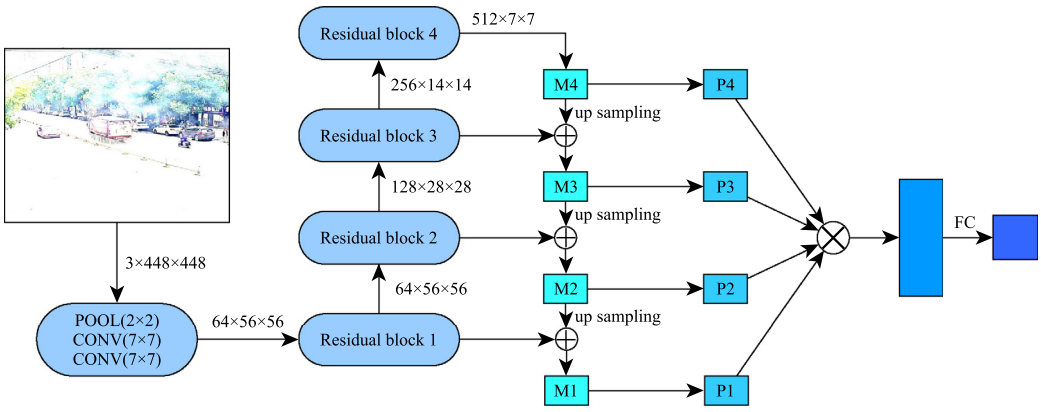

Fig. 2. Multi-scale feature fusion.

3.2. Multi-scale feature fusion

The multiscale feature fusion method can extract features at different scales, which is confirmed to be beneficial to image classification [37]. The usually used method of multiscale feature fusion is feature pyramid network (FPN) [38]. By fusing multiple scale feature maps together, receptive fields at different scales can be obtained, which can result in excellent performance on image classification.

In the task of surveillance video anomaly classification, the blurred anomaly detection is more difficult since deeper features lose the details of fuzzy and noise points. To better retain the feature information of fuzzy and noise points, we propose a multiscale feature fusion mechanism for surveillance video anomaly classification, which is shown in Fig. 2.

After residual block 4, we obtain a feature map with a size of 7 × 7. Then, we conduct a lateral connection process (1 × 1 convolution) and generate feature map M4 with the lowest resolution. By continuously upsampling M4, and combining it with the lateral connection feature map output from residual blocks 3–1, we obtain the feature maps of M3–M1 of different scales.

To eliminate the aliasing distortion effect caused by several times of upsampling, convolution is conducted on M4, M3, M2, and M1, and the features are smoothed to feature map sets P4, P3, P2, and P1. Four feature maps of different scales are concatenated together after a pooling operation. Finally, a full connection layer is used for classification with cross entropy loss, which computes the prediction probability values for each class. The formula for calculating multiclassification cross entropy is shown in Eq. (2). Cross entropy can evaluate the difference between the predicted probability distribution q and the real distribution p. The closer the probability of the two distributions is, the smaller the value of cross entropy represents and the better the prediction effect of the model is. Therefore, the cross entropy loss is very suitable for multiclass problems.

$$H(p, q) = -\sum_{x} p(x) \log q(x), \tag{2}$$

4. Experimental results and discussion

4.1. Datasets

We constructed a dataset for surveillance video diagnosis. The dataset contains 5 usual types of anomaly images. The 6 classes of 5 anomalies and 1 normal image are described as follows:

- **black–white**: the images captured by the color camera are black and white (not colorful).
- **blur**: the image is blurred and not distinct.
- **color_cast**: the color of the image does not match the color of the actual scene; for example, when the camera is reddish, white objects often appear red.
- **brightness**: the image is too dark or too bright.
- **occlusion**: the camera is occluded by sundries, such as leaves, plastic bag, etc.
- **normal**: the camera has no anomalies.

All the images are labeled by experts who have many years of experience in security and protection. For example, the camera occluded by leaves will be labeled "occlusion" whilst the camera occluded by large vehicles (such as bus) will not. The images are all collected from an actual surveillance video screen, and the size is 1920 × 1080. There are a total of 3349 images and approximately 500–600 images per class. Fig. 3 shows some samples of the dataset. We divide the total image samples into two parts, training and testing, at a ratio of 9:1.

4.2. Implementation and training details

Early parameter initialization methods generally initialize data and parameters to a standard Gaussian distribution, and the weight parameters W are selected as:

$$W \sim N\left(0, \frac{2}{n_i}\right), \tag{3}$$

where n_i indicates the ith layer's dimension. However, with increasing neural network depth, this method has difficulty solving the problem of vanishing gradients. In this paper, Kaiming initialization [39] is used as the parameter initialization method of the network since it is very suited for ReLU activation function. The images are resized to 448 × 448 and normalized before inputting for training. Random horizontal flipping is used for data augmentation. The batch size is set as 32. Exponential decay is used to adjust the learning rate, which is described as follows:

$$lr = lr \times \gamma^{e}, \tag{4}$$

where lr denotes the learning rate, e denotes the epoch number, and γ is set as 0.9. Our experiment is conducted with Python on an Intel i5-9400 CPU, Nvidia GTX-1660 GPU and 2 × 8Gbyte Memory PC without optimization.

4.3. Results and discussion

In this paper, the overall classification accuracy is used as the evaluation criterion for the multiclass task of surveillance video anomaly diagnosis. The higher the accuracy is, the better

Fig. 3. Examples of 5 types of anomalies and 1 normal image.

Table 2
Anomaly classification results by the proposed method.

Input type	Classified type					
	black-white	blur	color_cast	brightness	occlusion	normal
black-white	1.000	0.000	0.000	0.000	0.000	0.000
blur	0.000	0.964	0.000	0.000	0.000	0.036
color_cast	0.000	0.000	1.000	0.000	0.000	0.000
brightness	0.000	0.000	0.000	1.000	0.000	0.000
occlusion	0.000	0.020	0.000	0.000	0.980	0.000
normal	0.000	0.017	0.000	0.000	0.017	0.966

the ability of the technique to diagnose anomalies (e.g., brightness anomaly, blur anomaly, occlusion) is. The accuracy ratio is calculated as:

$$r = \frac{N_R}{N_T}, \tag{5}$$

where N_R represents the number of samples with correct classification, and N_T denotes the total number of samples in the testing dataset. The proposed technique archived an accuracy of 98.52%, which reveals the effectiveness of the framework.

Table 2 shows the classification details of each anomaly type. It gives the classification ratio of each anomaly class, which is defined as:

$$r_i^c = \frac{N_c}{N_i}, \tag{6}$$

Table 3
Comparison of the results with ResNet-18.

Method	Accuracy (%)
ResNet-18	92.88
Proposed	98.52

where N_i denotes the sample number of the input anomaly type, and N_c indicates the sample number of the classified anomaly type corresponding to the input type. From the results, we can see that each anomaly type can gain a high classification ratio, all of which are higher than 96%. This indicates that the proposed multiclass surveillance video anomaly diagnosis method indeed has the ability to detect and classify different anomalies.

To show the effectiveness of the proposed multiscale feature fusion mechanism, we conducted an ablation experiment that only uses the residual neural network ResNet-18. The comparison results are shown in Table 3, from which it can be seen that the proposed method is 6 percentage points higher than ResNet-18. Table 4 shows the classification details of each anomaly type by ResNet-18, which indicates that ResNet-18 cannot distinguish the blur, occlusion and normal types, although it can predict color_cast and brightness anomalies, just as the proposed method did. This is mainly because that the color_cast and brightness anomaly types have apparent global characteristics and are easier to recognize. However, the blur, occlusion and normal types have

Table 4
Anomaly classification results by ResNet-18.

Input type	Classified type					
	black-white	blur	color_cast	brightness	occlusion	normal
black-white	0.983	0.000	0.000	0.000	0.017	0.000
blur	0.000	0.909	0.000	0.000	0.000	0.091
color_cast	0.000	0.000	1.000	0.000	0.000	0.000
brightness	0.000	0.000	0.000	1.000	0.000	0.000
occlusion	0.000	0.020	0.000	0.000	0.939	0.041
normal	0.000	0.186	0.000	0.017	0.067	0.730

Table 5
Comparison of the results with state-of-the-art methods.

Method	Accuracy (%)	Time (s)
MANN [1]	93.47	0.0351
NNODSVM [32]	93.67	0.0410
Proposed	98.52	0.0379

more local differences among each other. The proposed multiscale feature fusion scheme can extract both global and local features; thus, the proposed method can distinguish all anomalies more effectively.

4.4. Comparisons with state-of-the-art methods

We compared the proposed method with some state-of-the-art methods: the morphological analysis and neural network (MANN) method [1], the neural network and optimization driven support vector machine (NNODSVM) method [32]. Table 5 shows the comparison results, which give the accuracy and average inference time. The accuracy of the proposed method is much higher than that of MANN [1] and NNODSVM [32]. The proposed method mainly benefits from the multiscale feature fusion scheme, which can extract both local and global features. Our method is also more effective in speed than NNODSVM, and it is fairly compared to MANN. The proposed technique reveals attractiveness for multiclass task of the diagnosis of different anomalies.

5. Conclusion

In this paper, a surveillance video diagnosis technique is proposed for multiclass anomaly classification. A multiscale feature fusion residual network is proposed to extract both the global and local features of the image. The global feature can predict anomalies that have “unitary” changes, such as color_cast and brightness. The local feature is capable of distinguishing anomalies which have “intricate” variations, such as blur and occlusion. Experimental results show that the proposed algorithm can gain a classification accuracy of 98.52%, which indicates the effectiveness of the proposed technique.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the NSFC, China fund (U2013210, 62006060, 62176077, 62002085), in part by National Key Research and Development Program of China under Project Number 2018AAA0100100, in part by the Guangdong Basic and Applied Basic Research Foundation, China under Grant 2019Bl515120055, 2021A1515012528, in part by the Shenzhen Key Technical Project, China under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund, China under Grant JCYJ20210324132210025, GXWD20201230155427003-20200824164357001, GXWD20201230155427003-20200824125730001 and in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China.

References

[1] L. Dong, Y. Zhang, C. Wen, H. Wu, Camera anomaly detection based on morphological analysis and deep learning, in: 2016 IEEE International Conference On Digital Signal Processing, DSP, IEEE, 2016, pp. 266–270.
[2] M.P. Ashby, The value of CCTV surveillance cameras as an investigative tool: An empirical analysis, Eur. J. Crim. Policy Res. 23 (3) (2017) 441–459.
[3] N. Li, X. Wu, H. Guo, D. Xu, Y. Ou, Y.-L. Chen, Anomaly detection in video surveillance via gaussian process, Int. J. Pattern Recognit. Artif. Intell. 29 (06) (2015) 1555011.
[4] N. Sulman, T. Sanocki, D. Goldgof, R. Kasturi, How effective is human video surveillance performance? in: 2008 19th International Conference On Pattern Recognition, IEEE, 2008, pp. 1–3.
[5] C. Kerdvibulvech, Human hand motion recognition using an extended particle filter, in: International Conference On Articulated Motion And Deformable Objects, Springer, 2014, pp. 71–80.
[6] C. Kerdvibulvech, Hybrid model of human hand motion for cybernetics application, in: 2014 IEEE International Conference On Systems, Man, And Cybernetics, SMC, IEEE, 2014, pp. 2367–2372.
[7] G.-F. Lu, Z. Jin, J. Zou, Face recognition using discriminant sparsity neighborhood preserving embedding, Knowl.-Based Syst. 31 (2012) 119–127.
[8] K. Dai, J. Zhao, F. Cao, A novel decorrelated neural network ensemble algorithm for face recognition, Knowl.-Based Syst. 89 (2015) 541–552.
[9] B. Chen, Z. Zhang, Y. Lu, F. Chen, G. Lu, D. Zhang, Semantic-interactive graph convolutional network for multilabel image recognition, IEEE Trans. Syst. Man Cybern. Syst. (2021).
[10] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[11] T. Özyer, D.S. Ak, R. Alhajj, Human action recognition approaches with video datasets—A survey, Knowl.-Based Syst. 222 (2021) 106995.
[12] X. Feng, W. Pei, Z. Jia, F. Chen, D. Zhang, G. Lu, Deep-masking generative network: A unified framework for background restoration from superimposed images, IEEE Trans. Image Process. 30 (2021) 4867–4882.
[13] L. Zhang, M. Luo, J. Liu, X. Chang, Y. Yang, A.G. Hauptmann, Deep top-k ranking for image–sentence matching, IEEE Trans. Multimedia 22 (3) (2019) 775–785.
[14] A. Sidnev, M. Barinova, S. Nosov, Efficient camera tampering detection with automatic parameter calibration, in: 2018 15th IEEE International Conference On Advanced Video And Signal Based Surveillance, AVSS, IEEE, 2018, pp. 1–6.
[15] A. Aksay, A. Temizel, A.E. Cetin, Camera tamper detection using wavelet analysis for video surveillance, in: 2007 IEEE Conference On Advanced Video And Signal Based Surveillance, IEEE, 2007, pp. 558–562.
[16] A. Saglam, A. Temizel, Real-time adaptive camera tamper detection for video surveillance, in: 2009 Sixth IEEE International Conference On Advanced Video And Signal Based Surveillance, IEEE, 2009, pp. 430–435.
[17] D.-Y. Huang, C.-H. Chen, T.-Y. Chen, W.-C. Hu, B.-C. Chen, Rapid detection of camera tampering and abnormal disturbance for video surveillance system, J. Vis. Commun. Image Represent. 25 (8) (2014) 1865–1877.
[18] M. Hagui, A. Boukhris, M.A. Mahjoub, Comparative study and enhancement of camera tampering detection algorithms, in: 2016 13th International Conference On Computer Graphics, Imaging And Visualization, CGiV, IEEE, 2016, pp. 226–231.
[19] Z. Liu, L. Meng, Y. Tan, J. Zhang, H. Zhang, Image compression based on octave convolution and semantic segmentation, Knowl.-Based Syst. 228 (2021) 107254.
[20] A. Levin, Blind motion deblurring using image statistics, Adv. Neural Inf. Process. Syst. 19 (2006) 841–848.
[21] H. Lee, C. Kim, Blurred image region detection and segmentation, in: 2014 IEEE International Conference On Image Processing, ICIP, IEEE, 2014, pp. 4427–4431.
[22] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, C. Zhang, Compound rank-k projections for bilinear analysis, IEEE Trans. Neural Netw. Learn. Syst. 27 (7) (2015) 1502–1513.
[23] S.A. Golestaneh, L.J. Karam, Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes, in: CVPR, 2017, pp. 596–605.
[24] T.A. Javaran, H. Hassanpour, V. Abolghasemi, Automatic estimation and segmentation of partial blur in natural images, Vis. Comput. 33 (2) (2017) 151–161.

[25] X. Yi, M. Eramian, Lbp-based segmentation of defocus blur, IEEE Trans. Image Process. 25 (4) (2016) 1626–1638.
[26] Y. Pang, H. Zhu, X. Li, X. Li, Classifying discriminative features for blur detection, IEEE Trans. Cybern. 46 (10) (2015) 2220–2227.
[27] J. Shi, L. Xu, J. Jia, Just noticeable defocus blur detection and estimation, in: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition, 2015, pp. 657–665.
[28] T. Kryjak, M. Komorkiewicz, M. Gorgon, FPGA implementation of camera tamper detection in real-time, in: Proceedings Of The 2012 Conference On Design And Architectures For Signal And Image Processing, IEEE, 2012, pp. 1–8.
[29] K. Sitara, B.M. Mehtre, Real-time automatic camera sabotage detection for surveillance systems, in: Advances In Signal Processing And Intelligent Recognition Systems, Springer, 2016, pp. 75–84.
[30] E. Ribnick, S. Atev, O. Masoud, N. Papanikolopoulos, R. Voyles, Real-time detection of camera tampering, in: 2006 IEEE International Conference On Video And Signal Based Surveillance, IEEE, 2006, p. 10.
[31] X. Yuan, A. Kortylewski, Y. Sun, A. Yuille, Robust instance segmentation through reasoning about multi-object occlusion, in: Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition, 2021, pp. 11141–11150.
[32] D. Liang, C. Lu, H. Jin, Soft multimedia anomaly detection based on neural network and optimization driven support vector machine, Multimedia Tools Appl. 78 (4) (2019) 4131–4154.
[33] X. Feng, H. Ji, B. Jiang, W. Pei, F. Chen, G. Lu, Contrastive feature decomposition for image reflection removal, in: 2021 IEEE International Conference On Multimedia And Expo, ICME, IEEE, 2021, pp. 1–6.
[34] X. Yuan, Z. Feng, M. Norton, X. Li, Generalized batch normalization: Towards accelerating deep neural networks, in: Proceedings Of The AAAI Conference On Artificial Intelligence, Vol. 33, 2019, pp. 1682–1689.
[35] K. Eckle, J. Schmidt-Hieber, A comparison of deep networks with ReLU activation function and linear spline-type methods, Neural Netw. 110 (2019) 232–242.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition, 2016, pp. 770–778.
[37] X. Fan, Y. Yang, C. Deng, J. Xu, X. Gao, Compressed multi-scale feature fusion network for single image super-resolution, Signal Process. 146 (2018) 50–60.
[38] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition, 2017, pp. 2117–2125.
[39] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings Of The IEEE International Conference On Computer Vision, 2015, pp. 1026–1034.