Pattern Recognition Letters 142 (2021) 20–24

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Violence detection and face recognition based on deep learning

Pin Wang [a], Peng Wang [b,*], En Fan [c]

[a] School of Mechanical and Electrical Engineering, Shenzhen Polytechnic, Shenzhen 518055, Guangdong, China
[b] Garden Center, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, Guangdong, China
[c] Department of Computer Science and Engineering, Shaoxing University, Shaoxing 312000, Zhejiang, China

ARTICLE INFO

Article history:
Received 2 May 2020
Revised 2 October 2020
Accepted 20 November 2020
Available online 16 December 2020

Keywords:
Deep learning
Brute force detection
Face recognition
Convolutional neural network
Video surveillance

ABSTRACT

With the emergence of the concept of "safe city", security construction has gradually been valued by various cities, and video surveillance technology has also been continuously developed and applied. However, as the functional requirements of actual applications become more and more diverse, video surveillance systems also need to be more intelligent. The purpose of this article is to study methods of brute force detection and face recognition based on deep learning. Aiming at the problem of abnormal behavior detection, especially the low efficiency and low accuracy of brute force detection, a brute force detection method based on the combination of convolutional neural network and trajectory is proposed. This method uses artificial features and depth features to extract the spatiotemporal features of the video through a convolutional neural network and combines them with the trajectory features. In view of the problem that face images in surveillance video cannot be accurately recognized due to low resolution, two models are proposed: the multi-foot input CNN model and the SPP-based CNN model. By testing the performance of the brute force detection method proposed in this paper, the accuracy of the method on the Crow and Hockey datasets is as high as 92% and 97.6%, respectively. Experimental results show that the violence detection method proposed in this paper improves the accuracy of violence detection in video.

1. Introduction

In recent years, the international situation has undergone tremendous changes, and violent terrorism and clashes of cliques have occurred frequently. In China, with the rapid development of Internet technology, video surveillance technology is also widely used, becoming the most important means of security surveillance. The success of deep learning has opened a new era of artificial intelligence, providing new solutions and solutions for intelligent video analysis technology. High-performance GPU devices solve the calculation problems caused by large-scale deep learning models and many parameters, and ensure timely event detection and processing in monitoring systems. At the same time, a large amount of video data also meets the requirements of a large number of data samples for deep learning training. This can effectively improve the system's data usage.

LeCun Y's team believes that deep learning allows computational models composed of multiple processing layers to learn data representations with multiple ion levels. Deep convolutional networks have made breakthroughs in image, video, speech, and audio processing, while recursive networks have inspired sequence data such as text and speech [1]. Marta Bautista-Durán's team believes that brute force detection is an important issue to consider when designing smart algorithms for smart environments. They proposed an energy-saving system that can detect violence in an acoustic manner. They found that the system is feasible [2]. Tim Valentine's team believes that the concept of multidimensional psychological space is the basis for modern theorists to deal with faces. They evaluated facial space as a theoretical framework for understanding ethnic influences and the development of facial recognition. And discussed two applications of facial space in the forensic environment [3].

2. Proposed method

2.1. Violence detection method

2.1.1. CNN model based on spatial features

Once the CNN model for extracting spatiotemporal features is trained, it will be used as a feature extractor to extract information between objects and scenes in video frames [4–5]. Assuming there is a test video V, then the convolutional feature layer that will be

* Corresponding author.
E-mail addresses: wangpin@vip.qq.com (P. Wang), sdhzdtwp@126.com (P. Wang), efan@szpt.edu.cn (E. Fan).

https://doi.org/10.1016/j.patrec.2020.11.018
0167-8655/

obtained is as follows:

$$C(V) = \{C_1^S, C_2^S, ..., C_M^S\} \tag{1}$$

$C_M^S \in Z^{H_m \times W_m \times L \times N_m}$is the feature layer of the $m \in \{1, 17\}$spatial network,$H_m$is the height of the feature layer,$W_m$is the width of the feature layer, L is the length of the video frame, and $N_m$is the number of convolution channels.

2.1.2. CNN model based on time series features

In order to obtain the timing information of the motion, this paper first extracts the optical flow field of the RGB image as the input of the CNN model according to the two-stream framework [6–7]. The optical flow field is the optical flow of a moving target, that is, velocity [8–9]. Optical flow records the difference between consecutive video frames and calculates the direction and speed of the target's movement, which can be used to describe the target's movement state [10–11]. In the video, the optical flow field can represent the motion change between two consecutive frames in a short time [12–13]. Once the model is trained, it will be used as a feature extractor to extract convolutional layers containing time-series features. Suppose there is a video V, then the convolutional feature layer that will be obtained is as follows:

$$C(V) = \{C_1^l, C_2^l, ..., C_M^l\} \tag{2}$$

$C_M^S \in Z^{H_m \times W_m \times L \times N_m}$is the feature layer of the$m \in \{1, 17\}$spatial network,$H_m$is the height of the feature layer, $W_m$is the width of the feature layer, L is the length of the video frame, and $N_m$is the number of convolution channels.

2.1.3. Trajectory extraction algorithm

Before tracking feature points, some feature points that are not key points need to be removed [14–15]. For example, the sampling points on the white wall on the background are not feature points. Here, the feature value of the autocorrelation matrix of each pixel is calculated, and a threshold is set to eliminate feature points below the threshold [16–17]. The threshold formula is determined by the following formula:

$$T = A \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2) \tag{3}$$

Where $(\lambda_i^1, \lambda_i^2)$is the characteristic value of the pixel and A is the adaptive coefficient. Assuming that the filtered feature point is $P_t = (x_t, y_t)$, then the position of the feature point in the next frame can be calculated using the following formula:

$$p_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{x_t, y_t} \tag{4}$$

Where $\omega_t = (u_t, v_t)$is the optical flow field, which is calculated from the current frame and the next frame, M is the median filter, $*$ is the convolution operation, and the size of the convolution kernel is 3*3. Therefore, by formula (4) to feature points are tracked in the optical flow field.

2.1.4. 3D trajectory depth feature

Suppose there is a video V, the trajectory $T_k$is obtained through the trajectory extraction, and the feature layer $C_m^n$is extracted through the depth feature extractor, where $n \in (s, t)$is the temporal feature map or the spatial feature map. This article uses the following formula to extract depth features based on trajectory:

$$F(T_k, C_m^n) = \sum_{l=1}^{L} (x_l^k \times ratio, y_l^k \times ratio, z_l^k) \tag{5}$$

Among them, $F(T_k, C_m^n)$is the depth-track feature extracted in this paper, and ratio is the scaling of the m-th feature layer.

2.2. Face recognition method

2.2.1. CNN deep network structure based on multi-scale input

In this paper, the convolutional neural network (CNN) in the deep model is selected for feature extraction. Using preprocessed labeled multi-resolution images, a multi-scale input convolutional neural network model (MRCNN) can be trained [18–19]. The input of the network is a mixed grayscale image with different resolutions, and the input size is 1 * 32 * 32. Two convolutional layers are connected to two maximum pooling layers to extract local information of the image [20–21]. The size of the convolution kernel is 5 × 5, the downsampling factor is set to 2, and the ReLU function is used as the nonlinear activation function. Through convolution operation, 64 5 × 5 feature maps can be obtained. These feature maps are vectorized and input into the fully connected layer fcl for further feature fusion. These features are also used as the final classification [22–23]. Finally, the features generated by fcl are classified by fully connected layers fc2 and LogSoftmax.

By minimizing the difference between the predicted value of the model and the corresponding label value, the parameter $\theta$of the model can be optimized. In this paper, the negative log-likelihood function is selected as the cost function of the model. For a given image set$X_i$and corresponding label informationY_i, the negative likelihood logarithm is defined as:

$$L(\theta) = \sum_{i}^{n} \sum_{k}^{K} -Y_i^k F(X_i; \theta)^k \tag{6}$$

Where K is the total number of categories, and n represents the number of samples in each category.

2.2.2. Deep network structure based on CNN and SPP

In this paper, a spatial pyramid pooling layer (SPP) is added before the fully connected layer of the CNN model. Through the SPP layer calculation, you can get maps of different sizes. By obtaining feature vectors of fixed length, the fully connected layer and even the entire model are guaranteed to work [24–25].According to the characteristics of the cross-space method, the cost function of low-resolution face recognition is defined as follows:

$$\min_{\theta} \alpha \sum_{i=j} dis(f(H_i, \theta) - f(L_j, \theta)) - \beta \sum_{i \neq j} dis(f(H_i, \theta) - f(L_j, \theta)) \tag{7}$$

Where $dis(\cdot)$is the distance function, f is the CNN model, θ is the model parameter, H and L correspond to high-resolution images and low-resolution images, respectively, and α and β are hyperparameters.

3. Experiments

3.1. Data collection

In order to verify the accuracy of brute force detection, this paper has conducted experiments on two public brute force data sets, namely Hockey data set and Crowd Violence data set. There are 500 videos containing violence in the hockey dataset. The videos of the crowd violence dataset are collected from real crowd scenes on YouTube. The data set has 140 videos with a resolution of 320 × 240 pixels and contains 123 non-violent videos. In order to verify the accuracy of the face recognition method, this paper selects 170 images with different expressions and lighting in the CMU PIE dataset for experiments. This paper also selected 2432 frontal images under all lighting conditions in the expanded Yale B dataset for experiments.

Table 1
Experimental results of different data sets.

	IDT	Spatial CNN	Time-lapse CNN	Spatial CNN+Time-lapse CNN+IDT
Crow	85	66	69	92
Hockey	90.1	90.5	86.3	97.6

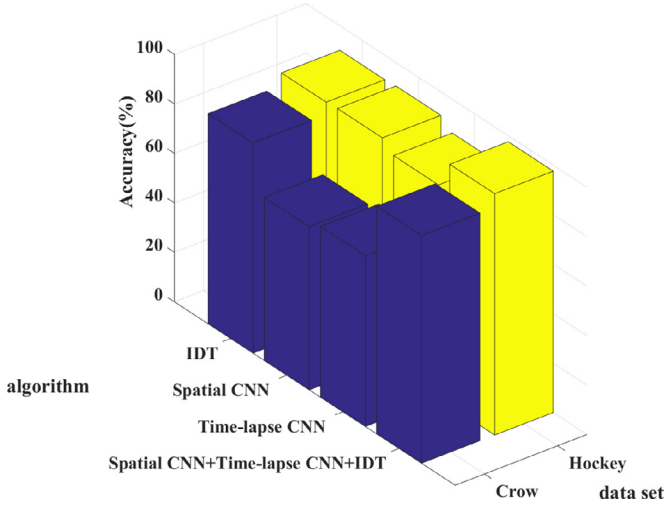

Fig. 1. Experimental results of different data sets.

3.2. Experimental setup

3.2.1. CNN model training and results based on spatial features

Training and results of CNN models based on spatial features. In this paper, the Hockey competition data set is selected as the training data. This article sets the training batch size to 50 and SGD to 0.9. Each video frame is re-modified to 340 × 256, and then 224 × 224 is randomly extracted from each frame as input data.

3.2.2. CNN model training and results based on time series features

Training and results of CNN models based on time series features. The input of the model is the optical flow field of the video frame. In this paper, the Farneback algorithm in the Opencv3.0 open source database is used to extract the dense optical flow field. The input data is 10 stacked optical flow fields. The network structure is the same as the CNN model based on spatial features. In this paper, 224 × 224 × 20 sub-regions of the input data are randomly selected. Here the learning rate is set to 0.005, and the training stops after 100 iterations.

4. Discussion

4.1. Analysis of experimental results of violence detection methods

4.1.1. Analysis of experimental results of different data sets

As can be seen from Table 1 and Fig. 1, as far as the Hockey data set is concerned, both the space and time series combined network model after trajectory extraction are better than the original space and time series CNN model. From the experimental results of Crowd database, the depth-trajectory feature extraction method in this paper performs better than other methods, and the best result of the experiment reaches 97.6%. This also shows that the method of combining manual features and deep features in the detection of violent behavior exceeds that of separate manual features or other deep network methods. The experimental results of different data sets are shown in Table 1 and Fig. 1.

4.1.2. ROC curve

The ROC curve uses the composition method to reveal the relationship between sensitivity and specificity. It is often used

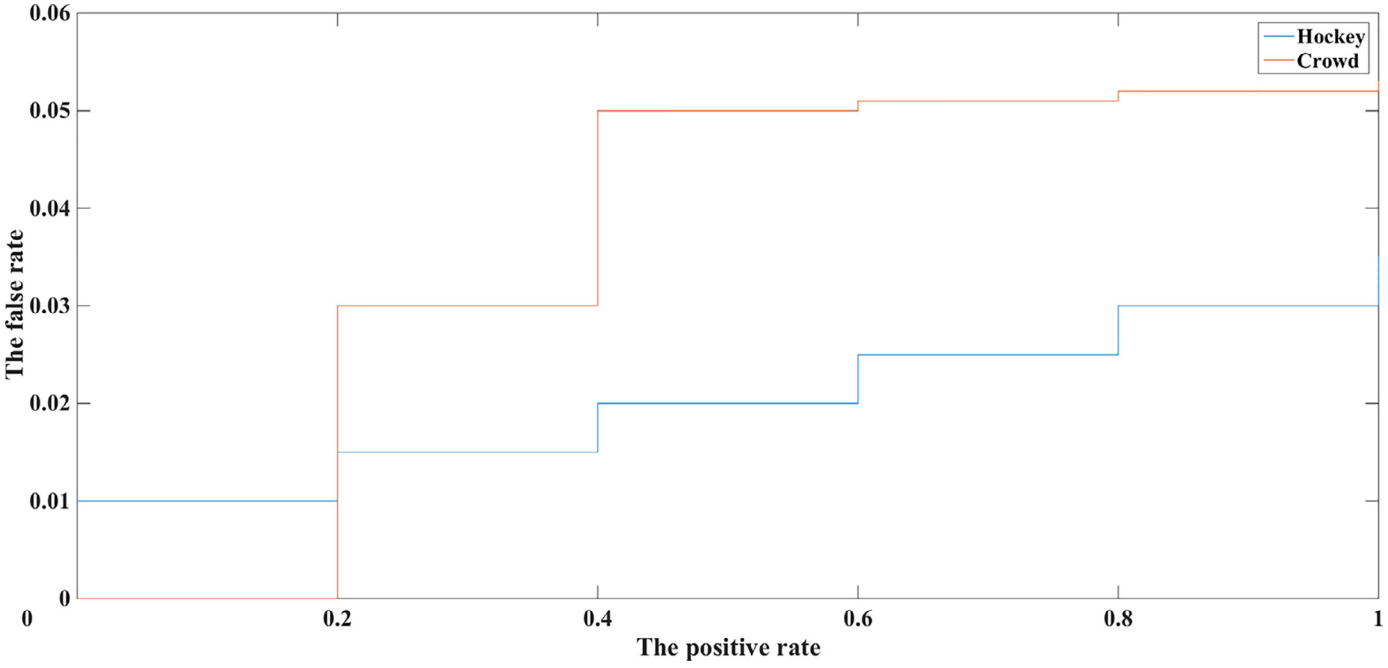

Fig. 2. ROC curve.

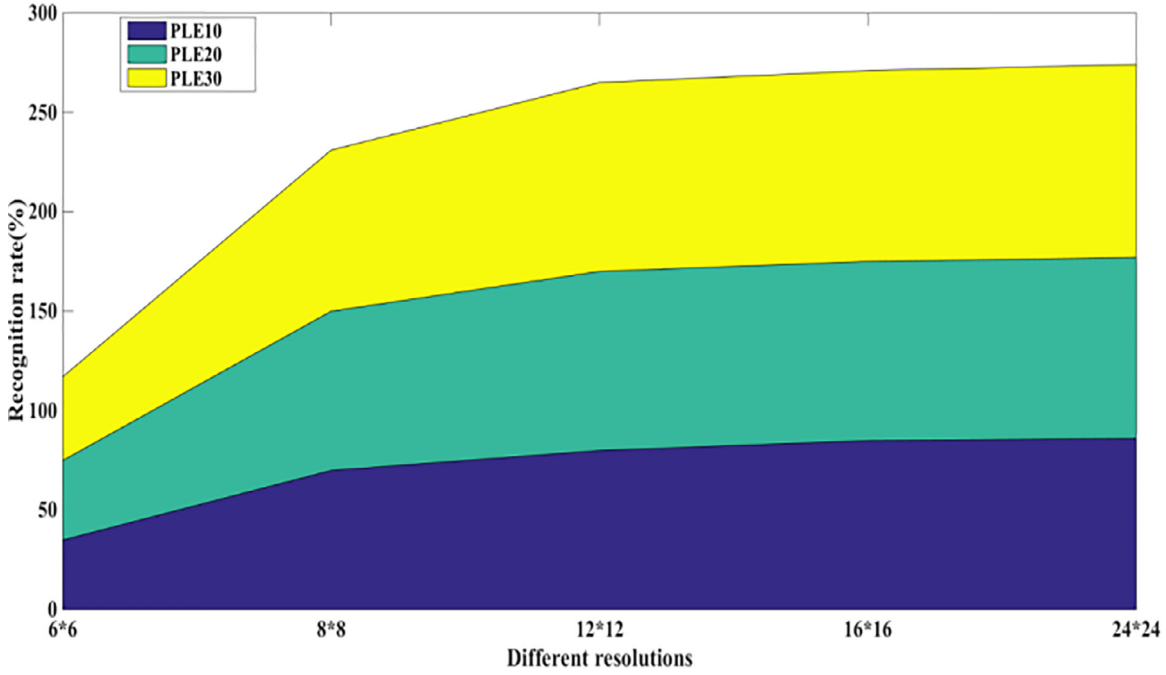

Fig. 3. Face recognition results at different resolutions.

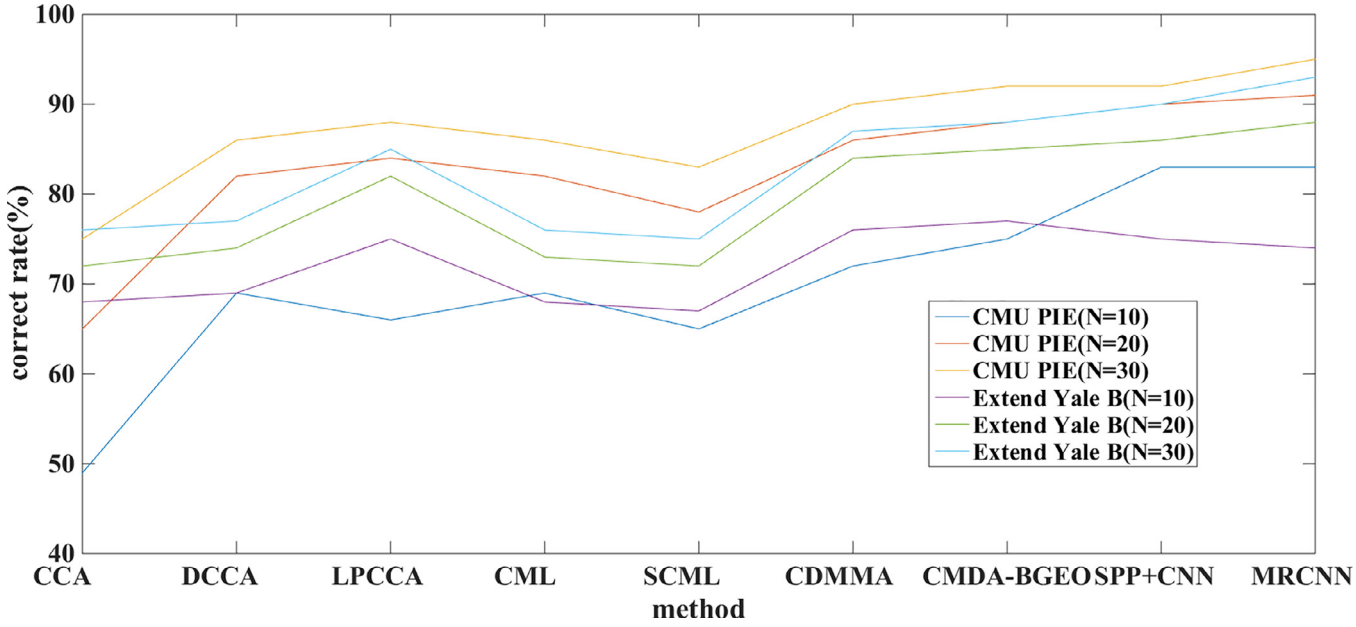

Fig. 4. Cumulative matching characteristic curve.

to evaluate the performance of a binary classifier. The closer the curve is to the lower right corner, the better the classifier performance. The results in Fig. 2 show that the method in this paper is very effective in detecting violent behaviors in videos, even in crowded crowd scenes. The ROC curve is shown in Fig. 2.

4.2. Analysis of experimental results of face recognition methods

4.2.1. CNN deep network structure based on multi-scale input

It can be seen from Table 2 and Fig. 3 that the MRCNN model achieved 86%, 91%, and 97% recognition accuracy in the case of training sample pairs at 10, 20, and 30, respectively. With the increase in the number of samples, the accuracy of the MRCNN model has been greatly improved, which also shows that the MRCNN model can achieve better results in the case of large amounts of data. The results of face recognition with different resolutions are shown in Table 2 and Fig. 3.

Table 2
Face recognition results at different resolutions.

	PLE10	PLE20	PLE30
6*6	35	40	42
8*8	70	80	81
12*12	80	90	95
16*16	85	90	96
24*24	86	91	97

4.2.2. Deep network structure based on CNN and SPP

As can be seen from Fig. 4, compared with the traditional cross-space method, the algorithm proposed in this paper achieves a better recognition effect under the condition of using only one projection matrix. Although compared with the MRCNN method, the model in this section has a slight performance gap, but the SPP-based CNN model eliminates the image preprocessing step, making the training process easier and faster, which is crucial in practical applications. The corresponding cumulative matching characteristic curve is shown in Fig. 4.

5. Conclusions

(1) Aiming at the problem of low efficiency and low accuracy of abnormal behavior detection in the monitoring system, this paper proposes a method of brute force detection based on the combination of CNN and ballistic. This method combines the advantages of manual features and deep learning features. By calculating the motion trajectory and the convolutional neural network feature layer, these two features are fused to obtain new features as criteria, thereby improving the accuracy of violent behavior detection.

(2) Aiming at the problem that face images in surveillance video cannot be accurately recognized due to low resolution, a low-resolution face recognition solution based on convolutional neural network model is proposed. The scheme proposed two models: multi-scale input continuous wave model and CNX model based on spp, CNN model based on spp, which belongs to the improved cross-space method.

(3) The recognition performance of the deep network model proposed in this paper for these data sets is still very limited. To solve this problem, on the one hand, we can improve the recognition rate and use a better model construction method through research. On the other hand, we can use the characteristics of video data to eliminate the influencing factors in the image, thereby further improving the accuracy of the model.

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 by National Natural Science Foundation of China under Grant 61703280.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning[J]., Nature 521 (7553) (2015) 436.
[2] M. Bautista-Durán, J. García-Gómez, R. Gil-Pita, Energy-efficient acoustic violence detector for smart cities, Int. J. Comput. Intel. Syst. 10 (1) (2017) 1298.
[3] T. Valentine, M.B. Lewis, P.J. Hills, Face-space: a unifying concept in face recognition research, Q. J. Exp. Psychol. 69 (10) (2016) 1996–2019.
[4] L. de Oliveira, M. Kagan, L. Mackey, Jet-images – deep learning edition, J. High Energy Phys. 2016 (7) (2015) 1–32.
[5] X. Lu, Z. Lin, H. Jin, Rating pictorial aesthetics using deep learning, IEEE Trans. Multimedia 17 (11) (2015) 1-1.
[6] X. Xiang Zhu, D. Tuia, L. Mou, Deep learning in remote sensing: a comprehensive review and list of resources, IEEE Geoence Remote Sensing Mag. 5 (4) (2017) 8–36.
[7] A. Kamilaris, F.X Prenafeta-Boldú, Deep learning in agriculture: a survey, Comput. Electron. Agric. 147 (1) (2018) 70–90.
[8] J. Han, D. Zhang, G. Cheng, et al., Advanced deep-learning techniques for salient and category-specific object detection: a survey, IEEE Signal Process. Mag. 35 (1) (2018) 84–100.
[9] D.B. Larson, M.C. Chen, M.P. Lungren, Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs, Radiology 287 (1) (2018) 313.
[10] O. Bernard, A. Lalande, C. Zotti, Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37 (11) (2018) 2514–2525.
[11] L. Mclaughlin, Improving intimate partner violence detection in the primary care setting: review of the literature, West. J. Nurs. Res. 39 (10) (2017) 1369.
[12] K. Solanki, P. Pittalia, Review of face recognition techniques, Int. J. Comput. Appl. 133 (12) (2016) 20–24.
[13] V.B. Nemirovskiy, A.K. Stoyanov, D.S. Goremykina, Face recognition based on the proximity measure clustering, Inst. Cybern. Tomsk Polytech. Univ. 40 (5) (2016) 740–745.
[14] E. Winarno, A. Harjoko, A.M. Arymurthy, Face recognition based on symmetrical half-join method using stereo vision camera, Int. J. Electr. Comput. Eng. 6 (6) (2016) 2818–2827.
[15] W. Zhe, G. Dan, F. Wei, Attention modulates the own-age bias in face recognition, Chin. Sci. Bull. 62 (15) (2017) 1620–1630.
[16] M.M. Kasar, D. Bhattacharyya, T.-H. Kim, Face recognition using neural network: a review, Int. J. Secur. Appl. 10 (3) (2016) 81–100.
[17] H.E. Khiyari, H. Wechsler, Face recognition across time lapse using convolutional neural networks, J. Inf. Secur. 07 (3) (2016) 141–151.
[18] S. Chakraborty, S.K. Singh, P. Chakraborty, Local directional gradient pattern: a local descriptor for face recognition, Multim. Tools Appl. 76 (1) (2017) 1201–1216.
[19] Z. Dong, Y. Wu, M. Pei, Vehicle type classification using unsupervised convolutional neural network, IEEE Trans. Intell. Transp. Syst. 16 (4) (2015) 1–10.
[20] F. Palsson, J.R Sveinsson, M.O. Ulfarsson, Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network, IEEE Geoence Remote Sensing Lett. 14 (5) (2017) 639–643.
[21] M. Hayder, T. Han, E. Naz, Hybrid algorithm for the optimization of training convolutional neural network, Int. J. Adv. Comput. Sci. Appl. 6 (10) (2015) 343–359.
[22] Y Xiang, Z. Lin, J. Meng, Automatic QRS complex detection using two-level convolutional neural network, BioMed. Eng. OnLine 17 (1) (2018) 13.
[23] X. Zhao, L. Liu, S. Qi, et al., Agile convolutional neural network for pulmonary nodule classification using CT images, Int. J. Comput. Assist. Radiol. Surg. 13 (1) (2018) 1–11.
[24] H.P. Trinh, M. Duranton, M. Paindavoine, Efficient data encoding for convolutional neural network application, ACM Trans. Architect. Code Optim. 11 (4) (2015) 1–21.
[25] C.P. Yu, T. Konkle, Map-CNN: a convolutional neural network with map-like organizations, J. Vis. 17 (10) (2017) 809.

Pin Wang was born in Dingtao County, Shandong Province, P.R. China, in 1983, she received the Ph.D. from Shenzhen University, P.R. China, in 2012. Now she is a teacher of Shenzhen Polytechnic. Her research interests include cloud data fusion, signal processing, and multi-target tracking.

Peng Wang was born in Dingtao County, Shandong Province, P.R. China, in 1987, He received the Master's degree from Zhongnan University of Economics and Law, P.R. China, in 2018. Now he is a staff of Garden Center, South China Botanical Garden, Chinese Academy of Sciences.

En Fan was born in Wuhan City, Hubei Province, P.R. China, in 1982. He received the Ph.D. degree from Xidian University, P.R. China, in 2015. Now he is a staff of Shaoxing University.