Real-time Face Recognition based on Pre-identification and Multi-scale Classification

Weidong Min [1], Mengdan Fan [1], Jing Li [1*], Qing Han [1]

[1] School of Information Engineering, Nanchang University, Nanchang, China
[*] corresponding. jingli@ncu.edu.cn

Abstract: In face recognition, searching a person's face in the whole picture is generally too time-consuming to ensure high detection accuracy. Objects similar to the human face or multi-view faces in low-resolution images may result in the failure of face recognition. To alleviate the above problems, a real-time face recognition method based on pre-identification and multi-scale classification is proposed in this paper. The face area is segmented based on the proportion of human faces in the pedestrian area to reduce the search range, and faces can be robustly detected in complicated scenarios such as heads moving frequently or with large angles. To accurately recognize small-scale faces, we propose the Multi-scale and Multi-channel Shallow Convolution network (MMSCN) which combines a multi-scale mechanism on the feature map with a multi-channel convolution network for real-time face recognition. It performs face matching only in the pre-identified face areas instead of the whole image, therefore it is more efficient. Experimental results showed that the proposed real-time face recognition method detects and recognizes faces correctly, and outperforms the existing methods in terms of effectiveness and efficiency.

Key words: Face recognition, Real-time systems, Face pre-identification, Deep learning, Multi-scale, Multi-channel

1. Introduction

With the goal of matching the capabilities of human vision [1], visual domain adaptation and generalization on static images or video frames have received significant attention in recent years. In particular, surveillance systems rely more and more on multi-view learning in the case that the source domain is different from the target domain in visual recognition [2]. Nevertheless, this task is relatively difficult and it requires real time and reliability.

Traditional face recognition methods recognize faces by detecting human faces, tracking the faces, capturing a frame of image, and then extracting features and matching faces in static images. For instance, Hamedani et al. [3] proposed a video-based face recognition method that first detected and tracked faces in videos, and then captured video frames of a subject rotating his/her head. The manifolds of video frames embedded in a high-dimensional video space were extracted using neural network-based models. After that, the images containing rotating heads were composited using nonlinear manifold learning and finally matched with the synthetic picture. On the other hand, some researches which integrate spatial and temporal probability models have been on the rise in recent years [4]. Tracking and identification in traditional face recognition algorithms are implemented separately, while in the spatial and temporal based method, they are performed at the same time, which is more real-time but a large amount of time information is needed both in the tracking and identification phases. In general, there are two main stages in the existing face recognition methods: i) detect human faces; ii) identify the detected faces. To detect human faces, it is necessary to search for the face in the whole image, and it may be misjudged when other objects are similar to the face. Furthermore, to ensure high detection accuracy, further filtering of the detected area is needed, which is time-consuming.

For reliability, some researchers focus on improving face detection methods, such as extracting hybrid features and improving classifiers [5]. However, when pedestrians are far away from the cameras, the target domain mainly consists of low-resolution images, while the source domain mainly contains high-resolution images. Moreover, the head moves more frequently than the body and recognizing multi-view faces becomes difficult when pedestrians look down, look up or look back. Other approaches improve face matching methods by using multi-scale properties of the facial features and multi-modal information [6], but require high-computational costs. In general, the existing methods still encounter the problem in low-resolution face recognition.

Overall, there are mainly two difficulties in real-time face recognition. One is that it is time-consuming to ensure high detection accuracy; the other is the difficulty to recognize faces in low-resolution images. In view of the first difficulty, we propose a pre-identification scheme, which combines pedestrian detection with face detection. To deal with the second difficulty, we propose a new deep learning network, the Multi-scale and Multi-channel Shallow Convolution network (MMSCN), for real-time face recognition. The MMSCN combines face pre-identification with multi-scale feature maps and multi-channel shallow convolution network, which not only improves the detection process of the traditional face recognition algorithms but also optimizes the recognition algorithms.

Firstly, in order to better capture small-scale target domain pictures, we combine pedestrian detection with face detection, and track pedestrians as the person pre-identification problem. The reason is that the basic principle of machine learning for detecting an object is to traverse the whole picture with a search box to find the most similar regions to the object. Large objects are more likely to be searched because there are less interfering pixels in the image [3]. Therefore, instead of searching faces directly, we obtain the area of the moving pedestrian by ACF features and SVM classifier, and then estimate the face area based on the proportion of human faces in the pedestrian area to reduce the search range of the detector. Afterwards, we estimate the face area according to the center of each pedestrian area. With the moving of pedestrians, a human face region selector is used to correct candidate face regions. This process is regarded as person pre-identification. Secondly, aimed at improving the face recognition performance on relatively small face images, the feature maps are down-sampled and central-clipped, respectively. Then, these two kinds of processed feature maps are trained separately by two network branches, and the features extracted from these two network branches are finally connected and used for classification.

The structure of this paper is as follows. Section 2 introduces the related works. In Section 3, we propose a novel face recognition method. Section 4 shows the experimental results and numerical simulations. Finally, conclusions and future work are given in Section 5.

2. Related Works

Originally, video-based face recognition algorithms identify human faces by the following steps: i) face detection; ii) face tracking; iii) feature extraction; and iv) face matching. Ren et al. [4] proposed a framework for dynamic scheduling for energy minimization (DSE) that leverages this emerging hardware heterogeneity. For instance, Schroff et al. [6] presented a system, called FaceNet, which directly learns a mapping from face images to a compact Euclidean space where distance directly corresponds to a measure of face similarity. To assess the performance of DSE, the authors built a face detection application based on the Viola-Jones classifier chain and conducted experimental studies via heterogeneous processor system emulation. On the other hand, several methods based on the integration of spatial and temporal probability models have been on the rise in recent years. For instance, Retter et al. [7] presented the human observers with natural images of objects at a fast periodic rate of 12.5 Hz, i.e., every 80 ms, uncovering that the neural spatio-temporal dynamics of category-selectivity in a rapid stream of natural images went well beyond previous evidence obtained from spatially and temporally isolated stimuli. Demirkus et al. [8] developed a fully automatic hierarchical and probabilistic framework that models the collective set of frame class distributions and spatial feature information over a video sequence. It is flexible enough to be applied to any facial classification tasks. Tracking and identification in the former method are performed separately, while in the latter they are performed at meanwhile. Therefore, the latter method is more efficient but a large amount of time information is needed both in the tracking and identification phases. Overall, the research of face recognition so far mainly focuses on the optimization of face recognition algorithms on static images, and the optimization of the spatial and temporal probability models of video-based face recognition.

Face recognition generally includes image capture, face detection, face pre-processing, feature extraction, and face classification. Luo et al. [5] proposed an adaptive skin detection method using face location and facial structure estimation. The face location algorithm was developed to improve the reliability of face detection and extracted a face region with a high proportion of skin. Common face recognition methods only use face detection to locate the position of the face to search human faces. When pedestrians are far from the cameras or in a complex background, it is difficult to track moving faces from video image frames. Several important enhancements made upon the original framework related to the pre-processing, feature calculation and training setup were described in [9]. On the other hand, pedestrian detection is widely used [10]. A pedestrian detection framework aided by information fusion between binocular visions was proposed in [11], which has the potential of aggregating information from multiple images to improve the detection on a single image. This work describes several important enhancements made in the original framework related to the pre-processing steps, feature calculation and training setup. Overall, advanced researches concerning face recognition are mostly based on machine learning, which finds the objects by traversing the whole image with a search box. Thus, large objects are more likely to be searched because there are less interfering pixels in the image [3].

Recently, deep learning-based algorithms have been widely applied in computer vision research area [3]. Wang et al. [12] proposed a joint method of priori convolutional neural networks at the superpixel level (called as “priori s-CNNs”) and soft restricted context transfer. It is worth noting that a priori s-CNNs model that learns priori location information at the superpixel level was proposed to describe various objects discriminatingly. A network trained end to end for better optimization which can be used for crowdedness regression tasks including congestion-level detection and crowd counting was proposed in [13]. Wang et al. [14] proposed a siamesed fully convolutional network (named as “s-FCN-loc”) based on the VGG-net architecture, which is able to consider the RGB-channel, semantic contour and location prior simultaneously to segment the road region elaborately. Although the deep learning methods mentioned above have achieved great success in those computer vision tasks, the network structures are complex to a certain extent, and not all deep learning networks are suitable for face recognition problems. It is worth thinking how to make the network structure simpler.

In order to improve the face recognition rate, multi-view face detection [13] has been widely proposed in face recognition. In the video-based face recognition process, with the uncertainty of pedestrian movement in the target area, the face is likely to appear with different sizes, so it is necessary to normalize the size of face images. But if the size of the captured human face from a video is relatively small, the obtained face image needs to be enlarged, which reduces image resolution. To solve this problem, a bottom-up saliency detection method that unites the syncretic merits of sparse representation and multi-hierarchical layers was proposed. In contrast to most pre-existing sparse-based approaches that only highlight the boundaries of a target, the method highlights the entire object even if it is large [15]. He et al. [16] equipped the networks with spatial pyramid pooling. The network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. Pyramid pooling is also robust to object deformations. Although the recognition performance is improved, the detection algorithm in the above methods remains the same. Searching a person's face in the whole picture is time-consuming which requires complex calculation and could bring in the delay of the algorithm. In contrast, a computationally efficient face recognition method based on multi-scale images could be a good substitute. Overall, it is not easy to recognize low-resolution faces for the existing face recognition methods.

To this end, this paper aims to alleviate the issues discussed above, which are low face detection and recognition rate, and high time consumption. The contributions of this paper are as follows:

1) We perform face detection based on pre-identification, in which the face area is captured based on the proportion of human faces in the pedestrian area. It could reduce the search range, and the faces can be robustly detected in complicated scenarios such as heads moving frequently or with large angles.

2) We propose a novel deep learning network MMSCN to recognize faces in multiple views by combining multi-scale feature maps with a multi-channel shallow convolution network.

3. Proposed Method

The key steps of the real-time video-based face recognition are as follows: 1) face detection: detect and mark the faces in video frames; 2) face tracking: regard the coordinates of detected human faces as the initial state of the tracking target, and then track the faces; 3) feature extraction: extract the face features such as principal components from the entire face; 4) feature dimension reduction: reduce the extracted high-dimensional features to low-dimensional ones; 5) face recognition: perform face classification based on face features.

On the basis of the above contents, we propose a real-time face recognition method based on pre-identification detection and multi-scale classification.

3.1. Proposed Pre-identification detection

3.1.1. Overall process: The flowchart of the whole algorithm is depicted in Fig. 1. Firstly, the ACF features are extracted and classified by SVM. Secondly, NMS is used to improve the algorithm by getting multiple overlapped borders (bounding box) and reducing them to only one border. Thirdly, the face area is estimated based on the proportion of the pedestrian area to reduce the search range, which is regarded as pre-identification.

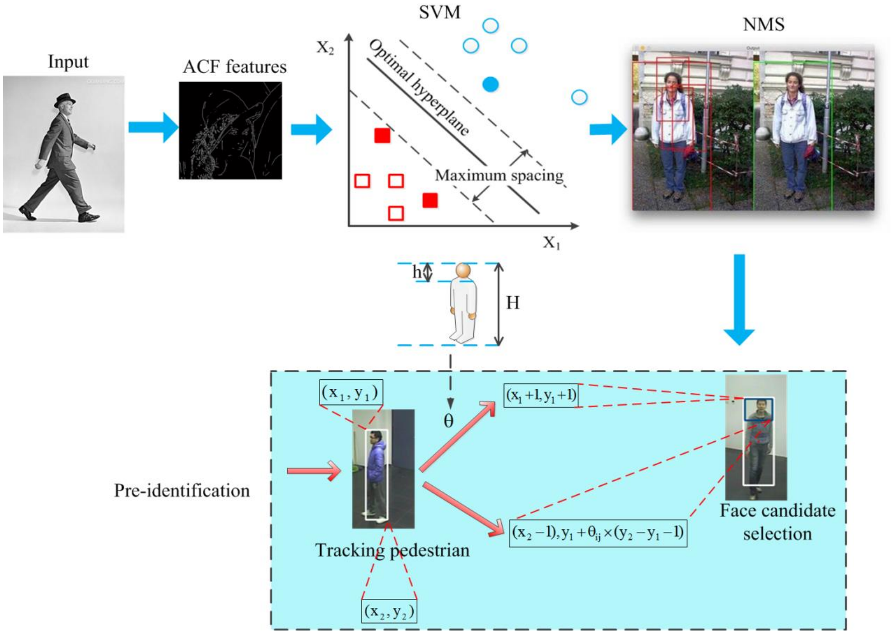


Fig. 1. *Detection based on pre-identification.*

3.1.2. Principle of pre-identification detection: According to previous sections, general face recognition methods search human faces throughout the whole image. When pedestrians are far from the cameras or in a complex background, it is hard to track moving faces from video sequences. In the real-time application scenarios, the prior information of the scene can be used to determine the potential area of the face, which reduces the search range of the detector. Considering this, we obtain the moving pedestrian area by a pedestrian detection algorithm and estimate the face area based on the proportion of the pedestrian area to reduce the search range, which is referred as pre-identification.

Pedestrian detection involves the method based on background modelling and the method based on classifiers. Background modelling can be divided into background subtraction and frame differencing. Background subtraction detects moving objects by comparing the current frame with the background reference model through the determination of the change of gray level [17-19], or use the histogram and other statistical information to segment the moving object [20-22]. It is one of the most commonly-used methods of moving-object detection in video sequences. Frame differencing [23, 24] is a method to obtain the target contour through differential operation of adjacent video frames. For fast-moving objects, smaller intervals need to be selected; otherwise the objects in two adjacent frames without overlap will be detected as two separate objects. For slow-moving objects, the larger time difference should be chosen; otherwise the objects in two adjacent frames that completely overlap will not be detected.

The pedestrian detection methods based on classifiers are divided into AdaBoost, SVM and deep learning methods. For instances, the cascade implementation of the additive KSVM (AKSVM) was proposed in [25]. AKSVM avoids kernel expansion by using lookup tables, and it is implemented in a cascade form, thereby speeding up pedestrian detection. The cascade implementation is trained by a genetic algorithm such that the computation time is minimized, whereas the detection accuracy is maximized. Wu et al. [26] proposed a semi-supervised approach for training deep convolutional networks on partially labelled data. It is hypothesized that the components of the auxiliary detector capture essential human characteristics that are useful for constructing a scene-adapted detector. Dominguez-Sanchez et al. [27] proposed a CNN-based technique that leverages current pedestrian detection techniques, which are used as an input for the proposed modified versions of various state-of-the-art CNN networks, such as AlexNet, GoogleNet, and ResNet. The detection methods based on deep learning are better in recognition performance. In summary, the detection efficiency of the derivatives of these methods is basically the same [28], but SVM is more concise than the AdaBoost, and the number of parameters is less than that of CNN.

According to the methods mentioned above, we detect the moving pedestrian area by ACF features [29] and SVM classifier [30], and then use the non-max suppression (NMS) [31] to remove redundant frames. Finally, the face area based on the proportion of the pedestrian area is estimated to reduce the search range, which is regarded as pre-identification. The entire algorithm will be described in Section 3.1.4. In order to further verify the effectiveness of the pre-identification scheme, we compare it with the detection process in advanced face detection algorithms in terms of accuracy and time efficiency in Section 4.

3.1.2 Pedestrian detection: Based on the comprehensive comparison of Section 3.1.1, we combine ACF and SVM to detect moving objects. Firstly, the ACF features are extracted and classified by SVM. Secondly, non-max suppression (NMS) is used to improve the algorithm by getting multiple overlapped borders (bounding box) and reducing them to only one border.

ACF is one of the most advanced feature extraction algorithms [29]. The ACF features are robust by superimposing the LUV color, the gradient amplitude and the gradient direction features. SVM can be applied to solving other problems, but in this section it is mainly used based on the core idea introduced in [30]. The NMS method is mainly used to eliminate redundant frames through searching for local maxima and suppress non maxima elements through getting rid of the redundant box based on the classifier score for each candidate [31].

3.1.3 Pre-identification mechanism: The human face area is estimated according to the proportion of the pedestrian area to reduce the search range, which is regarded as pre-identification. With the moving of the pedestrians, a face-region selector is used to correct face candidate regions, which is described as follows. The whole process of face pre-identification before face detection in each image frame is shown in Fig. 1. Our face detection algorithm based on pre-identification is described totally in Algorithm 1.

Algorithm 1 Face detection

```
Input: Image or video
Use Depth-2 decision tree as the weak classifier and soft-cascade structure to extract feature $z$ .
The training samples are located on both sides of the hyperplane.
if ( $z$ is linear separable)
    A total $T = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)...\}$ samples, the hyperplane is $w \cdot x + b = 0$ , the geometric distance from the sample point to the hyperplane is: $\frac{y_i \cdot y(x_i)}{\|w\|} = \frac{y_i \cdot (w^T \cdot \Phi(x_i) + b)}{\|w\|}$;
else
    Look for a point on the horizontal axis and calculate the function value of this point. It is a time curve we know, and its function expression can be written as: $g(x) = c_0 + c_1 x + c_2 x^2$ .
Let vector $y$ and $a$ be as the follows: $a = [a_1, a_2, a_3]^T = [c_0, c_1, c_2]^T$ $y = [y_1, y_2, y_3]^T = [1, x, x^2]^T$ then we have: $g(x) = ay$
Left-most candidate boxes of the input sequence stand for the candidate boxes.
if ( $I[j] > I[j+1]$ and $I[j] \geq I[j-1]$ )
    $MaximumAt(i)$;
else $i$ is not an local maximum
```

```
12: Divide the whole image into $p \times q$ regions, where each region corresponds to a selector $\theta$: Where $H$ is the height of the pedestrian and $h$ is the height of the human face;
13: Get the upper-left corner of the pedestrian area and the lower-right corner through the target detection and tracking algorithms;
14: Calculate the center of each target in the region $\alpha_{ij}$;
15: Estimate human face regions by the following two formulas: The upper-left corner of the area which may contain human faces is: $(x_1+1, y_1+1)$; The lower-right corner of the area which may contain human faces is: $((x_2-1), y_1+\theta_{ij}\times(y_2-y_1-1))$
```

3.2. Proposed Multi-scale classification

Before recognizing the face, we pre-process face images to improve the image quality by eliminating the irrelevant information in the image so as to enhance the detection ability. Pre-process includes geometric transformation (rotate and translate the image) and normalization (clip the image), histogram equalization (improve brightness and contrast), smoothing (reduce noise). After that, we propose a new deep learning network MMSCN for real-time face recognition which combines multi-scale feature maps and multi-channel shallow convolution network.

Different from static images, it is not easy to determine how pedestrians move in the target area in video sequences. Since face images may have substantial changes, normalization should be considered. Here, images of small sizes are enlarged, and the resolution is reduced which could lead to wrong recognition results. In order to solve the above problems, we propose the MMSCN, as shown in Fig. 2. It consists of two network branches. The feature maps are down-sampled and central-clipped, respectively. Then, these two kinds of processed feature maps are trained separately by two network branches, and the features extracted from these two network branches are finally connected and used for classification.

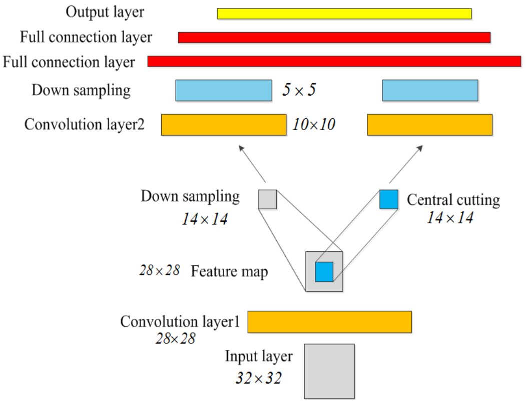


***Fig. 2.** Our proposed MMSCN*

Network: Fig. 2 shows the MMSCN. In order to better recognize the face at different scales, inspired by [32], we combine down sampling and central cutting. Down sampling can reduce the resolution of the feature map and prepare for recognizing small-scale images with low resolution. The resolution of the original feature map is maintained after central cutting, but the scale has also been declined. To reduce the computational complexity, the patch is cut out from the feature map instead of the original image. At the same time, the network structure is made into a dual channel mode in order to simplify the parameter setting.

Multi-scale processing: Here, the feature maps are processed in multiple scales. Firstly, the original image is used to search out the candidate window of face. Secondly, the entire image to be identified is entered to CNN, feature maps are obtained, and then pooled by different methods, i.e., down sampling and central cutting. The idea of multi-scale processing is to split the original feature map of 64*64 into two 32*32 sub-maps. The first sub-map is directly obtained through down sampling of the original feature map. The second sub-map is obtained by using central cutting to cut out a 32*32 sub-map in the center of the origin feature map, which is shown as the light blue area in Fig. 2.
Source domain and Target domain: $D=\{x, P(X)\}$, D represents the domain, x represents the feature space, and $X = x_1, x_2, .. x_n \in x$. $P(x)$ represents the marginal probability distribution, and x_i is the binary feature of the i -th pixel block .

Learning tasks: A task T can be made up of a tag space Y and an objective function $f_T(\cdot)$. The objective function can be expressed as a form of conditional probability distribution $P(Y|x)$. For two categories, it is either true or false.

Learning goals: The source domain D_s is integrated with the features of different-scale images. The target domain D_t mainly contains small-scale pictures. The purpose is to improve the classification result of the objective function $f_T(\cdot)$ in D_t .

4. Experimental and Numerical Simulations

The experimental environment used in the experiments is: Intel Celeron i5-2400 @3.10GHz CPU, 4GB internal storage, Windows 7 32bit operating system, Axis 215 PTZ network camera, 30 frames per second as the frame rate for network cameras.

Since our algorithm is based on pedestrian detection, to demonstrate the superiority of our algorithm, the training samples of the database must contain both the whole body and the face of the pedestrians. Due to the lack of such a dataset, we collected a dataset containing 5 different subjects. For each subject, 900 frames are sampled. Besides, we made a standard dataset by fusing the data from the processed LFW dataset [33] and the processed PIE dataset [33] together since they are similar in type to some extent. In the processed LFW dataset, there are more than 13,000 pictures in total, containing 5,765 different identities.

The PIE dataset contains 750,000 face images from 337 different identities with different attitudes, expressions and illumination changes. In order to better reflect the superiority of the proposed algorithm on multi-scale face recognition, the images in the test set of the standard dataset are sampled to 28×28 . Then, the images at this scale will be normalized to the input size before being tested by the model. Padding after down-sampling greatly reduces the resolution of the original map, and thus can be used to verify whether low-resolution data can be accurately detected.

4.1. Face detection efficiency with pre-identification

With the moving of pedestrians, the detection method uses human face region selector to correct face candidate regions, as described in section 3.1. The correction operator obtains faces regions through the segmented regions, and dynamically adjusts the detection window in real-time to ensure the detection effect. Fig. 3 shows the candidate face regions obtained by the correction operator at different moments on the self-collected dataset. As we can see, no matter what it is, e.g., a side face, far-distance face, near-distance face or face with partial-body unseen, the pre-identification method could detect the face accurately.

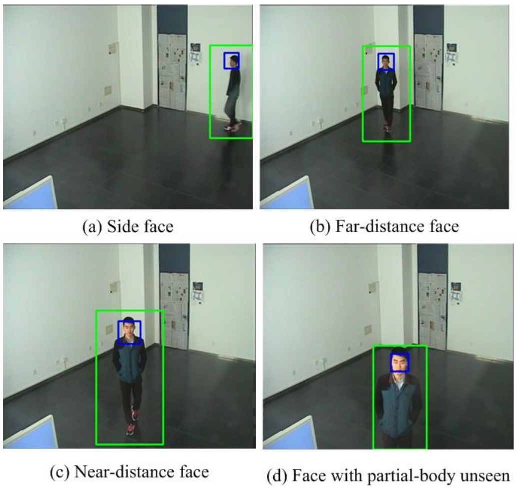

(a) Side face
(b) Far-distance face
(c) Near-distance face
(d) Face with partial-body unseen

Fig. 3. *Candidate face regions obtained in different moments*

To demonstrate the superiority of pre-identification, we compare the detection results with two of the most advanced face detection algorithms, which are MTCNN [6] and Viola-Jones [4] by qualitative and quantitative analysis. In particular, because the standard dataset does not contain the whole body of pedestrians, we implement qualitative analysis only on our newly collected dataset.

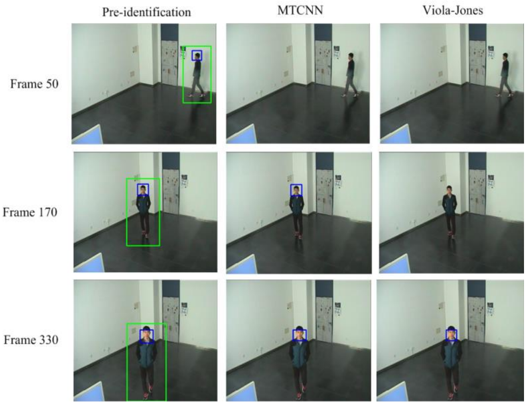


Fig. 4. *Qualitative analysis*

In Fig. 4, it can be seen that both MTCNN and Viola-Jones cannot filter out the background when the face is fairly small, which may lead to misjudge. On the contrary, the algorithm proposed in this paper can solve this problem very well. The results of the quantitative analysis are shown in Fig. 5 and Table 1. Because of the lack of a standard face detection dataset that contains the whole body, we use the ROC curve and fps as criteria to evaluate the algorithms based on the self-collected dataset (Video frame rate is 30 fps). Fig. 5 shows the ROC curves of the proposed detection algorithm and the advanced detection algorithms on the self-collected dataset. It can be seen that pre-identification achieves better performance.

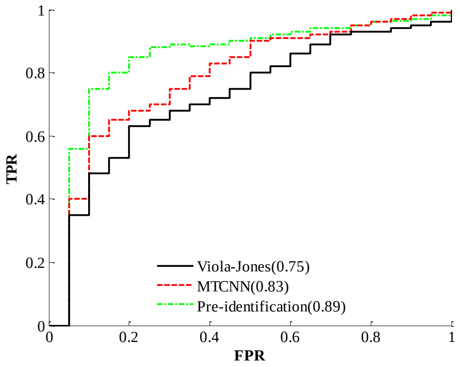


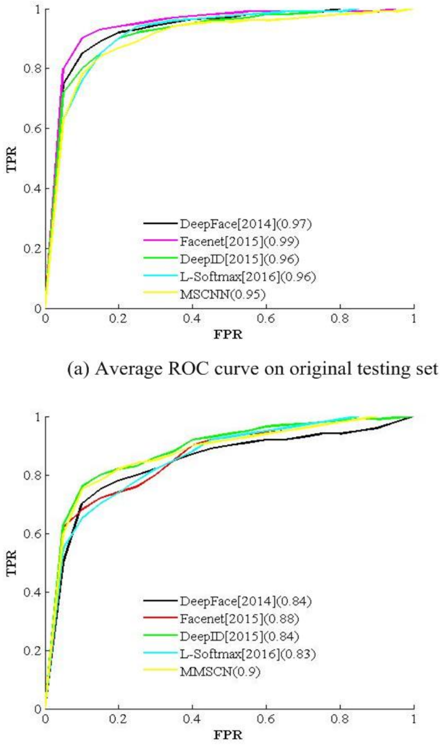
Fig. 5. *ROC curves of the proposed detection algorithm and the advanced detection algorithms on the self-collected dataset.*

Table 1 Comparison of time consumption

Face detection algorithm	Frame rate (fps)
V-J	27.509
MTCNN	26.748
Pre-identification scheme	29.779

4.2. Recognition effect of MMSCN

In order to verify the effectiveness of the MMSCN, we conduct qualitative and quantitative comparisons with some of the state-of-the-arts face recognition methods on processed LFW dataset [33] and processed PIE dataset [33]. The training set accounted for seventy percent, and the testing set was thirty percent. In order to better reflect the superiority of the proposed algorithm on multi-scale face recognition, the images in the testing set of the standard dataset are sampled to 28×28. Moreover, the image at this scale will be normalized to the input size before being tested in the models. This greatly reduces the resolution of the original map. Fig. 6 shows the comparison results on the original testing set and the small-scale testing set.



(a) Average ROC curve on original testing set

(b) Average ROC curve on Small-scale testing set

Fig. 6. *Average ROC curves of recognition algorithms on the standard dataset.*

As we can see from Fig. 6, although FaceNet obtains better classification performance on the original testing set, the MMSCN performs better on the small-scale testing set. The greater the area of the area under curve (AUC) is, the better the performance of the algorithm. Fig. 6 proves that our algorithm has higher recognition rate on small-scale images.

4.3. Overall recognition effect

In order to test the robustness of the whole algorithm, we compare the recognition results of the four state-of-the-arts face recognition methods. They are DeepFace [34], FaceNet [6], DeepID [35], and L-Softmax [36]. The results are shown in Fig. 7. It can be seen that the proposed method can successfully and accurately detect the face once it just appeared in the video. As seen from Fig. 7, although the FaceNet algorithm has obtained the highest recognition accuracy so far [6], its detection method is not robust enough. Moreover, its structure is relatively complex, which requires a large amount of computation. In summary, the algorithm proposed in this paper can accurately detect faces, and it obtains a higher recognition rate compared with advanced methods. Moreover, the model is simple and suitable for real-time face recognition systems.

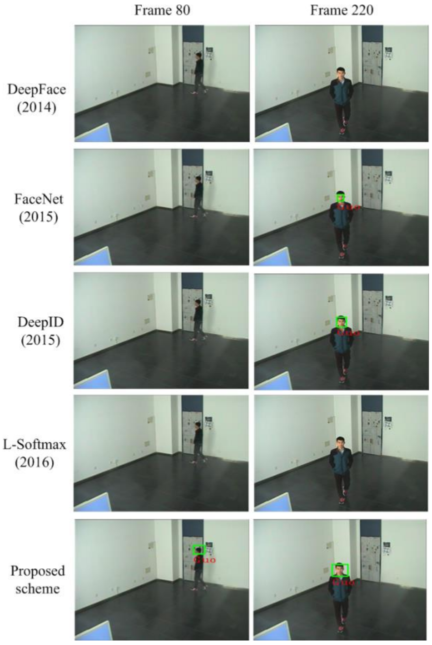


Fig. 7. *Qualitative results of recognition*

5. Conclusion

A real-time face recognition method based on pre-identification and multi-scale classification is proposed in this paper. Firstly, we combine pedestrian detection with face detection, and track pedestrians as person pre-identification. After obtaining the area of the moving pedestrian by ACF features and SVM, we estimate the face area based on the proportion of the pedestrian area to reduce the search range of the detector. Secondly, aimed at improving the face recognition performance on relatively small face images, we propose Multi-scale and Multi-channel Shallow Convolution network (MMSCN) to recognize the face. Experimental results showed that the proposed face recognition method outperforms the existing methods in terms of effectiveness and efficiency.

There still exists several open studies in the above research. In regarding of the pre-identification method, the potential studies will be oriented to studying better mechanisms for predicting the range of faces. Furthermore, the depth of the MMSCN structure could be increased. We have only made a preliminary study with limited experiments that prove the effectiveness of the strategy.

6. Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant 61463032, 61703198, 61762061 and the Natural Science Foundation of Jiangxi Province, China under Grant 20161ACB20004.

7. References

[1] Bondi, L., Baroffio, L., Cesana, M., Tagliasacchi, M., Chiachia, G., Rocha, A.: 'Rate-energy-accuracy optimization of convolutional architectures for face recognition', J vis commun image represent, 2016, 36, (1), pp. 142–148.

[2] De-la-Torre, M., Granger, E., Sabourin, R., Gorodnichy, D.O.: 'Adaptive skew-sensitive ensembles for face recognition in video surveillance', Patt Recog, 2015, 48, (11), pp. 3385–3406.

[3] Hamedani, K., Seyyedsalehi, S. A., Ahamdi R.: 'Video-based face recognition and image synthesis from rotating head frames using nonlinear manifold learning by neural networks', Neural Comput and Applic. 2015, 27, (6), pp. 1–9.

[4] Ren, S., Deligiannis, N., Andreopoulos, Y., Islam, M., Schaar, M.: 'Dynamic Scheduling for Energy Minimization in Delay-Sensitive Stream Mining', IEEE Trans Signal Process., 2014, 62, (20), pp. 5439–5448.

[5] Luo, Y., Guan, Y. P.: 'Adaptive skin detection using face location and facial structure estimation', Iet Computer Vision., 2017, 11, (7), pp. 550–559.

[6] Schroff, F., Kalenichenko, D., Philbin, J.: 'FaceNet: A unified embedding for face recognition and clustering', Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), MA, USA, 2015, pp. 815– 823.

[7] Retter, T. L., Rossion, B.: 'Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream', Neuropsychologia., 2016, 91, (1), pp. 9–28.

[8] Demirkus, M., Precup, D., Clark J. J., Member S., Arbel T.: 'Hierarchical Spatio-Temporal Probabilistic Graphical Model with Multiple Feature Fusion for Binary Facial Attribute Classification in Real-World Face Videos', IEEE Trans. Pattern Anal. Mach. Intell., 2016, 38, (6), pp. 1185 – 1203.

[9] Bilal M.: 'Algorithmic optimisation of histogram intersection kernel support vector machine-based pedestrian detection using low complexity features', Iet Computer Vision., 2017, 11, (5), pp. 350 – 357.

[10] Mazzon, R., Tahir, S., Cavallaro A.: 'Person re-identification in crowd', Pattern recognition let., 2012, 33, (14), pp. 1828 – 1837.

[11] Ren H., Li, Z.: 'Object detection using boosted local binaries', Pattern Recognition., 2016, 60, (1), pp. 793 – 801.

[12] Wang, Q., Gao, J., Yuan, Y.: 'A Joint Convolutional Neural Networks and Context Transfer for Street Scenes Labeling', IEEE Trans. Intell. Transp. Syst., 2018, 19, (5), pp. 1457 – 1470.

[13] Wang, Q., Wan, J., Yuan, Y.: 'Deep Metric Learning for Crowdedness Regression', IEEE Trans. Circ Syst. Video T., 2017, pp.

[14] Gao, J., Wang, Q., Yuan, Y.: 'Embedding structured contour and location prior in siamesed fully convolutional networks for road detection' Proc. Int. Conf. Robotics and Automation, Singapore, Singapore, May 2017, pp. 219–224.

[15] Hu, Z., Zhang, Z., Sun, Z., Zhao S.: 'Salient object detection via sparse representation and multi-layer contour zooming', Iet Computer Vision., 2017, 11, (4), pp. 309 – 318.

[16] He, K., Zhang, X., Ren, S., Sun. J.: 'Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition ', IEEE Trans. Pattern Anal. Mach. Intell., 2014, 37, (9), pp. 1904 –1916.

[17] Sato, Y. D., Nagatomi, T., Horio, K., H.: 'Miyamoto, The Cognitive Mechanisms of Multi-scale perception for the recognition of extremely similar faces', Cogn. Comput., 2015, 7, (5), pp. 501 – 508.

[18] Liu, X., Zhao, G., Yao, J., Qi S, 'Background subtraction based on low-rank and structured sparse decomposition', IEEE Trans. Image. Processing., 2015, 24, (8), pp. 2502 – 2514.

[19] Xie, Y., Yang, L., Sun, X., Wua, D., Chen, Q, Yong, M., 'An auto-adaptive background subtraction method for Raman spectra', Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy., 2016, 161, (1), pp.58 – 63.

[20] Ge, W., Guo, Z., Dong, Y., Chen Y, 'Dynamic background estimation and complementary learning for pixel-wise foreground/background segmentation', Pattern Recognition., 2016, 59, (1), pp. 112 – 125.

[21] Huang, M., Chen, Y., Ji, W., Miao C., 'Accurate and Robust Moving-Object Segmentation for Telepresence Systems', ACM Trans. Intell. Syst. Tech., 2015, 6, (2), pp. 1 – 17.

[22] Liang, C. W., Juang, C. F., 'Moving Object Classification Using a Combination of Static Appearance Features and Spatial and Temporal Entropy Values of Optical Flows', IEEE Trans. Intell. Transpsyst., 2015, 16, (6) , pp. 3453 – 3464.

[23] Zhang, Y., Lu, H., Zhang, L., Ruan, X., 'Combining motion and appearance cues for anomaly detection', Pattern Recognition., 2016, 51, (1), pp. 443 – 452.

[24] Shi, X., Shan, Z., Zhao, N., 'Learning for an aesthetic model for estimating the traffic state in the traffic video', Neurocomputing., 2016, 18, (1) , pp. 29 – 37.

[25] Baek, J., Kim, J., Kim, E., 'Fast and Efficient Pedestrian Detection via the Cascade Implementation of an Additive Kernel Support Vector Machine', IEEE Trans. Intell. Transpsyst., 2017, 18, (4), pp. 902 – 916.

[26] Wu, S., Wang, S., Laganiere, R., Liu, C., Wong, H. S., Xu, Y., 'Exploiting Target Data to Learn Deep Convolutional Networks for Scene-Adapted Human Detection', IEEE Trans. Image. Processing., 2018, 27, (3), pp. 1418 – 1432.

[27] Dominguez-Sanchez, A., Cazorla, M., 'Orts-Escolano S. Pedestrian Movement Direction Recognition Using Convolutional Neural Networks', IEEE Trans. Intell. Transp. Syst., 2017, 18, (12), pp. 3540 – 3548.

[28] Htike, K. K., Hogg, D., 'Adapting pedestrian detectors to new domains: A comprehensive review', Eng. Applartif. intell., 2016, 50, (1), pp. 142–158.

[29] Yang, B., Yan, J., Lei, Z., Li, S. Z.: 'Aggregate channel features for multi-view face detection', Proc. Int. Conf. Biometrics (IGCB), Clearwater, FL, USA 2014, pp. 1– 8.

[30] Wu, J., Yang, H.: 'Linear Regression-Based Efficient SVM Learning for Large-Scale Classification', IEEE Trans. Neural. Netw. Learn. Syst. 2017, 26, (10), pp. 2357 – 2369.

[31] Song, Y., Glasbey, C. A., Polder, G.: 'Non-destructive automatic leaf area measurements by combining stereo and time-of-flight images', Iet Computer Vision., 2014, 8, (5), pp. 391 – 403.

[32] Zagoruyko, S., Komodakis, N., 'Learning to Compare Image Patches via Convolutional Neural Networks', in Int. Conf. Computer Vision and Pattern Recognition (CVPR). MA, USA, 2015, pp. 4353 – 4361.

[33] Soltanpour, S., Boufama, B., Wu, Q., 'A Survey of Local Feature Methods for 3D Face Recognition', Pattern Recognition, 2017, pp. 72–87.

[34] Taigman Y, Yang M, Ranzato M, Wolf L., 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification', Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, June 2014, pp. 1701-1708.

[35] Sun Y, Liang D, Wang X, Tang X., 'DeepID3: Face Recognition with Very Deep Neural Networks', Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), MA, USA, 2015, pp. 1-5.

[36] Liu W, Wen Y, Yu Z, and Yang M., 'Large-margin softmax loss for convolutional neural networks', Proc. Int. Conf. Machine Learning (ICML), 2016, pp. 507-516.

Author names and affiliations

Weidong Min

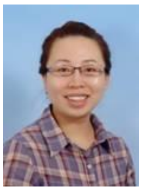
He obtained his BE, ME and PhD of computer application at Tsinghua University in China in 1989, 1991 and 1995, respectively, on the research subjects of computer graphics, image processing and computer aided geometric design. He was an assistant professor of Tsinghua University from 1994 to 1995. From 1995 to 1997 he was a postdoctoral researcher at University of Alberta, Canada. From 1998 to 2014 he worked as a senior researcher and senior project manager at Corel and other companies in Canada. In recent years, he cooperated with School of Computer Science & Software Engineering, Tianjin Polytechnic University, China. From 2015 he is a professor at School of Information Engineering, Nanchang University, China. He is a Member of “The Recruitment Program of Global Expert” of Chinese government. He is an executive director of China Society of Image and Graphics. His current research interests include computer graphics, image and video processing, distributed system, software engineering and network management.

Mengdan Fan

She obtained her BE of computer application at Jiangxi Agricultural University in China in 2015. She is a postgraduate at Nanchang University in China now, on the research subject of abnormal behavior detection in video surveillance.

Jing Li

Jing Li obtained her PhD degree in Electronic and Electrical Engineering from the University of Sheffield, UK, in 2012. Before joining Nanchang University as an Associate Professor, she was a Research Associate at the University of Sheffield. Her research interests include content-based image retrieval, object recognition, visual tracking and scene understanding in complex environments. She has authored or co-authored in various journals, such as IEEE Transactions on Industrial Informatics, Information Sciences (Elsevier), etc.



Qing Han

She obtained her BE and ME of computer application at Tianjin Polytechnic University in China in 1997 and 2006, respectively. She is now an associate professor at School of Information Engineering, Nanchang University, China. Her research interests include image and video processing, network management.