ICIC Express Letters
Volume **17**, Number **4**, April **2023**

 ISSN 1881-803X
pp. **449–456**

VIDEO FACE RECOGNITION COMBINING WITH MULTI-SCALE FEATURES AND FRAME STRUCTURE PERCEPTION

Ning Ouyang, Hewei Zhang* and Leping Lin

School of Information and Communication
Guilin University of Electronic Technology
No. 1, Jinji Road, Guilin 541004, P. R. China
{ ouyangning; linleping }@guet.edu.cn
*Corresponding author: ynou@guet.edu.cn

Received July 2022; accepted September 2022

Abstract. *Because the images in the face video clips are affected by factors such as occlusion and light changes, the quality of each frame is poor, and the actual video sequence is too long, which leads to the complexity and low accuracy of the video face recognition model. In order to solve the above problems, a frame structure-aware aggregation network is proposed to construct an overall video frame feature representation. First, the multi-scale feature extraction module is used to learn the feature representation of video frames, and then the feature aggregation network is trained and the corresponding weights are assigned to the feature representation of each video frame to achieve the purpose of evaluating the importance of frames. At the same time, the context information is effectively modeled by combining the mining of frame relationships. Compared with the traditional method of selecting key frames for recognition, the method in this paper can more efficiently utilize the features of each video frame and its spatial structure information. The results of experiments on two public video face recognition datasets, IJB-A and YTF, show that our scheme has a certain improvement in video face recognition performance.*

Keywords: Video face recognition technology, Multi-scale feature extraction module, Feature aggregation network, Correlation between frames

1. **Introduction.** A breakthrough in face recognition technology will improve the efficiency of tasks such as video surveillance and person identification [1]. Video has one more time dimension than images, which can be understood as an ordered set of images. It is particularly important of how to extract feature representations from video sequences that are beneficial to recognition. At present, there are mainly two types of advanced video face recognition models: 1) extracting key frames for recognition, 2) using all frames of video sequences for recognition. We believe that low-quality frames still have important value for the preservation of video integrity and structural information. [2] proposes a new aggregation adversarial network, which aggregates high-quality face images from low-quality video frames through the competitive relationship formed by the aggregation module and the discriminator. [3] proposes a context-aware feature aggregation scheme to perceive complementary information between video frames. [4] proposes a video face recognition algorithm based on aggregated local spatiotemporal descriptors, which aggregates temporal and spatial continuity information in videos. Based on this, this paper performs feature fusion on all video frames, so that it keeps the filtering of noise information and fuses the inter-frame structure information.

Methods for face recognition using whole video sequences are mainly divided into two stages, namely learning more accurate features and building frame aggregation models. [5] proposes a deep aggregation model (Neural Aggregation Network, NAN), which creatively

DOI: 10.24507/icicel.17.04.449

separates the feature extraction and aggregation stages. Descriptors assign weights and then fuse the feature representations of all video frames together by weighted averaging, but NAN does not exploit the temporal dimension of the video. On this basis, [6] proposes a recurrent embedded aggregation neural network, in the process of feature aggregation by redesigning the long short-term memory to integrate temporal information, thereby avoiding noise introduced by redundant video frames. [7] proposes an ordered weighted aggregation network that introduces an ordered weighted average operator in the process of video frame feature aggregation, which can combine the feature information of each element with the sorting, reducing the model while improving the recognition accuracy.

The second part of this paper mainly introduces our two important contributions, the third part will be divided into three subsections to describe the overall framework of this scheme, the fourth part will design experiments to verify the scheme of this paper, and the final summary and outlook are in the fifth part.

2. **Our Main Contributions.** The scheme in this paper considers the complex interrelationships and location structure information between video frames, and efficiently fuses features to improve the performance of video face recognition models. This paper proposes an inter-frame structure-aware aggregation network combined with multi-scale feature extraction for video face recognition. The main contributions are as follows.

1) Design a multi-scale feature extraction network to learn video frame feature representations. The feature extraction network designed in this paper maintains efficient learning of features at different scales while adapting to local feature scale changes.

2) A frame structure-aware aggregation network is proposed to aggregate the features of video frames. Considering the mutual competition and cooperation between each video frame, the features are aggregated more efficiently.

3. **Overall Framework.** Figure 1 shows the overall framework of the scheme in this paper, which is mainly divided into two stages: 1) In the feature extraction stage, input n frames of video frames to be tested, and then use the designed multi-scale feature extraction network to learn the feature representation $\{\boldsymbol{f}_n\}$ of each frame, so as to obtain the overall feature sequence of the video; 2) The feature aggregation stage receives the output of the previous stage as input, which is used to train the frame structure-aware

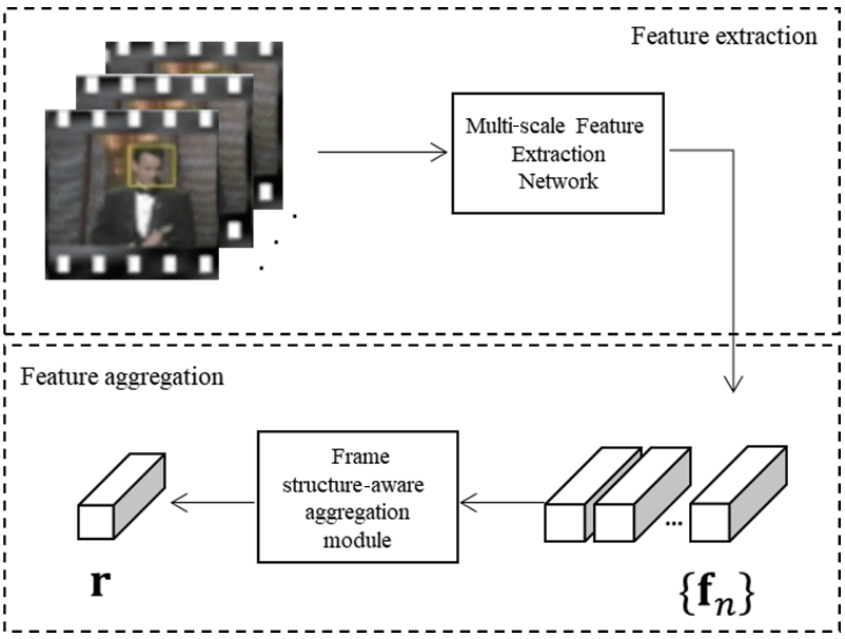

FIGURE 1. The overall framework of the program

aggregation network to adaptively predict the weights for each frame feature vector. Finally, these sequences, which fuse the overall information of the video and the structural information between frames, are aggregated into a feature vector $\boldsymbol{r}$. This feature vector will be used to complete subsequent recognition tasks.

3.1. **Multi-scale feature extraction network.** The multi-scale feature extraction network designed in this paper is shown in Figure 2. The actual number of convolutional layers will be appropriately deepened, in order to establish a convolutional neural network with strong semantics on multiple scales. Taking this feature, we introduce this structure into the feature learning process of video faces to extract features of different scales of faces, and also adapt to the scale changes of local features caused by video frames over time. This also greatly improves the efficiency of the subsequent feature aggregation process.

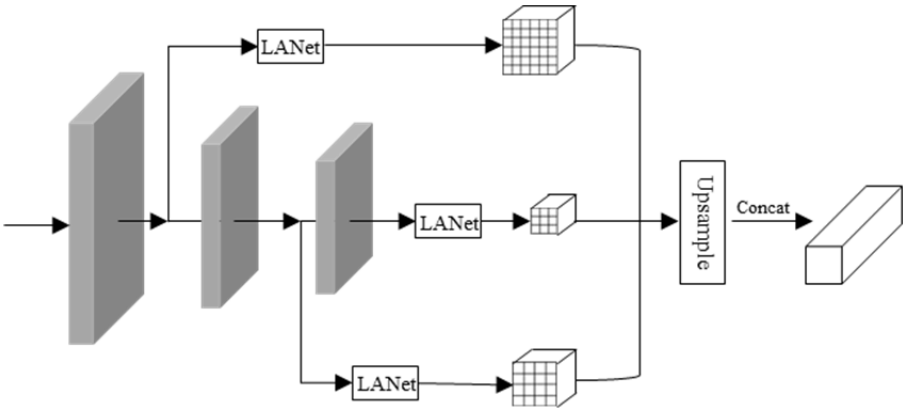

FIGURE 2. Multi-scale feature extraction network

The network can be divided into three parts, namely global feature extraction, local feature extraction and feature fusion operations. The backbone convolutional network gradually reduces the input feature map scale, while the Local Aggregation Network (LANet) is used to adaptively learn local descriptors. After obtaining the feature maps of multiple scales and local details of the face, deconvolution is used for upsampling, and finally the Concat fusion operation is performed.

LANet is based on the framework given in [8]. This scheme inserts LANet between the backbone convolutional layers, which consists of two convolutional layers of size, which are used to aggregate inter-channel spatial information into one channel. The application of this structure makes more informative local features have higher attention, while less important features will be further ignored.

3.2. **Frame structure-aware aggregation module.** As an individual, a single video frame has unique feature representation and location information. On the other hand, when we look at the video as a whole, each frame will have a certain correlation with other frames, and it also has an important impact on the integrity of the video. Consider a video clip $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \ldots, \boldsymbol{f}_n\}$, where $\boldsymbol{f}_i$ is the feature vector of the video frame, and n represents the number of frames in the video. The details of the frame structure-aware aggregation module proposed in this paper are shown in Figure 3.

Defining $f_i : f_j$ as the mutual relationship between the i frame and the j frame, this relationship is calculated by the following formula:

$$s_{i,j} = f_i : f_j = \phi_s(\boldsymbol{f}_i)^T \psi_s(\boldsymbol{f}_j) \tag{1}$$

where ϕ_s and ψ_s are called embedding functions. In the process of training the network to mine the structural relationship between frames, the 1×1 convolution and BN layers

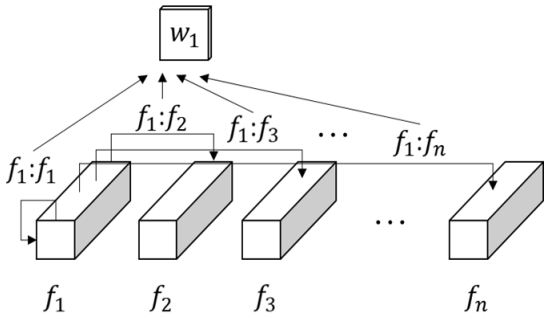

FIGURE 3. Frame structure-aware aggregation module

are used to realize these two operations, and the ReLU activation function is introduced to obtain

$$\phi_s(\boldsymbol{f}_i) = \mathrm{ReLU}(W_\phi \boldsymbol{f}_i) \tag{2}$$

$$\psi_s(\boldsymbol{f}_j) = \mathrm{ReLU}(W_\psi \boldsymbol{f}_j) \tag{3}$$

Then, use $\boldsymbol{S}_{(i,:)} = [s_{i1}, s_{i2}, s_{i3}, \ldots, s_{in}]$ to represent the relationship vector between the ith frame and each video frame. In contrast, $\boldsymbol{S}_{(:,i)} = [s_{1i}, s_{2i}, s_{3i}, \ldots, s_{ni}]$ is used to represent the relationship vector between each video frame and the ith frame, where $i = 1, \ldots, n$. These two vectors fuse the position and structure information of the frame corresponding to the video segment $\boldsymbol{F}$. Finally, we combine $\boldsymbol{S}_{(i,:)}$ and $\boldsymbol{S}_{(:,i)}$ by the following formula:

$$\sigma_i = \mathrm{Sigmoid}\left(W_2 \mathrm{ReLU}\left(W_1 \boldsymbol{S}_{(i,:)}^T \boldsymbol{S}_{(:,i)}\right)\right) \tag{4}$$

where σ_i is called the position factor or the structure factor, and this parameter can better help us determine the importance of a certain frame in the video sequence. We introduce this factor when subsequently training the network to assign weights to video frames.

We denote the descriptor obtained after the fusion of relation vector and frame feature as $\boldsymbol{v}_i$, where $i = 1, 2, \ldots, n$. The descriptor includes not only all the features of each frame, but also its structural relationship and position information with the overall video frame. Next, flatten $\boldsymbol{v}_i$ into a D-dimensional vector to obtain the corresponding weights by

$$e_n = \sigma_n \boldsymbol{q}^T \boldsymbol{f}_n \tag{5}$$

$$w_n = \frac{\exp(e_n)}{\sum_i \exp(e_i)} \tag{6}$$

The structure factor assigns weights based on structure information to each frame while the initialization kernel $\boldsymbol{q}$ is multiplied by the frame features, and the final video face feature $\boldsymbol{r}$ is aggregated by the following formula:

$$\boldsymbol{r} = \sum_{i=1}^{n} w_i \boldsymbol{v}_i \tag{7}$$

By combining context and video frame structure information, the less important frames in the fusion process will further reduce the influence, while also making the resulting feature vectors more discriminative. Figure 4 presents the results of assigning weights to typical examples in the dataset using our method.

3.3. **Training of the network.** During the training of the model, we combine the two modules, resulting in an end-to-end training approach. First, do not introduce σ_n and initialize $\boldsymbol{q}$ with all-zero, and train on the experimental dataset; then fix the parameter $\boldsymbol{q}$, then introduce σ_n and further fine-tune the model. The average contrastive loss [9] is minimized by constructing two shared-weight frame structure-aware aggregation networks, while the optimizer uses Adam [10].

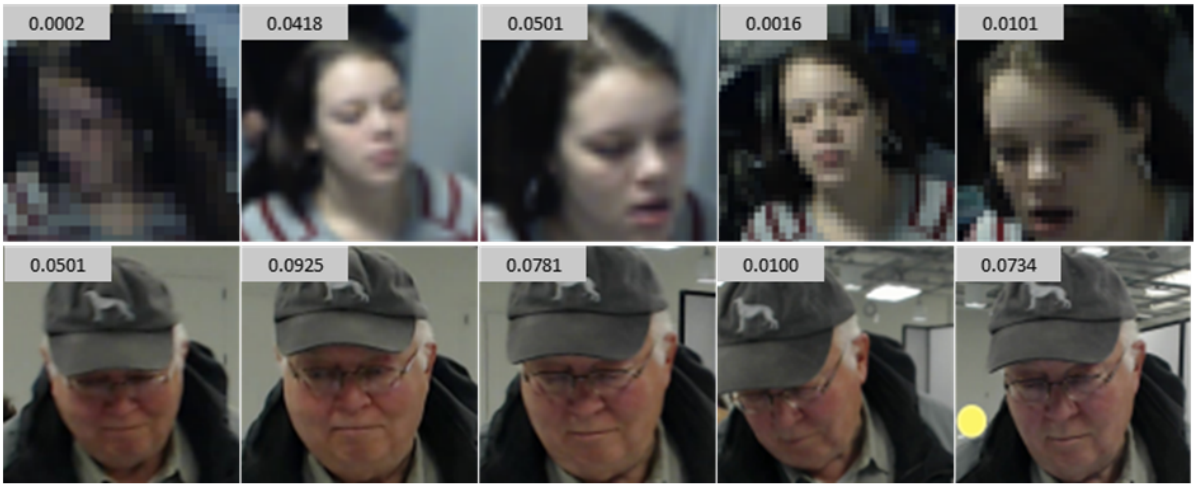

FIGURE 4. Typical examples showing the weights of the images in the image sets computed by our method

4. Experimental Results and Analysis.

4.1. Dataset of the experiment. To evaluate the proposed scheme, we conduct experiments on two publicly available video face recognition datasets: IJB-A [11] and YTF [12]. IJB-A is a dataset for face detection and recognition. It contains 55,026 video frames and 7,011 video clips, most of which were shot in unconstrained environments in the wild, which are very close to the faces captured by video surveillance data. The YTF dataset contains 3,425 videos of 1,595 objects, which are from users on the YouTube website and most of them use mobile devices to shoot, so the quality of face videos is slightly higher.

In the preprocessing stage, the MTCNN model [13] is used to detect face images in the dataset, and the input video frame size is 224×224. In experiments, we compare this scheme with simple aggregation strategies such as average pooling and 2D convolution, and also with state-of-the-art video face recognition models.

The experimental environment is based on Nvidia GeForce RTX 2080Ti GPU, the initial learning rate is set to 0.001, the batch size is set to 128, and the number of training iterations is set to 500.

4.2. Experimental evaluation based on dataset IJB-A. In the experimental evaluation process for the IJB-A dataset, video clips of 384 objects were selected as input, and TAR and FAR were introduced to evaluate the performance of each model. FAR means False Accept Rate, which is equal to the proportion of treating objects with different labels as the same one. TAR means True Accept Rate, which is equal to the proportion of objects whose similarity is greater than the threshold T when comparing objects with the same label. Under the same FAR, the higher the TAR value, the better the model performance. Table 1 shows the experimental results of different models on this dataset.

TABLE 1. Comparison of TAR@FAR using IJB-A dataset

Methods	Verification TAR (%)		
	FAR = 0.001	FAR = 0.01	FAR = 0.1
NAN	87.66	95.72	97.69
DCNNs [14]	–	83.80	96.70
CNN+AvePool	84.14	93.54	96.87
VGGFace [15]	80.51	91.50	95.33
Ours	**87.73**	**96.32**	**98.34**

In the experiment, the comparison model CNN+AvePool simply average pools the features to generate the final overall feature representation of the video frame. NAN uses weighted average to give weights to each video frame in the process of feature aggregation and then combine them. DCNNs are unconstrained face image recognition models based on deep convolutional networks. VGGFace is a traditional static face recognition model.

As can be seen from Table 1, DCNNs, CNN+AvePool and VGGFace perform the worst among all the models in the experiments. Compared with the latter two, NAN has a 2-4 percent higher TAR when FAR = 0.01. The model proposed in this paper has a TAR of 87.73% and 96.32% when the FAR is 0.001 and 0.01, respectively. When FAR = 0.1, our method also outperforms the comparison model. In general, the method in this paper has the best performance among the four models in the comparative experiments.

In order to discuss the influence of the number of input video frames on the method in this paper, we also changed the number of frames to test accordingly, and the results are shown in Figure 5. As can be seen from the figure, when the number of input frames is 15 frames, the experimental results of the method in this paper based on the IJB-A dataset achieve the best results. When FAR = 0.01, TAR = 0.9632. However, when the number of input frames exceeds 15 frames and gradually increases, the model performance gradually decreases.

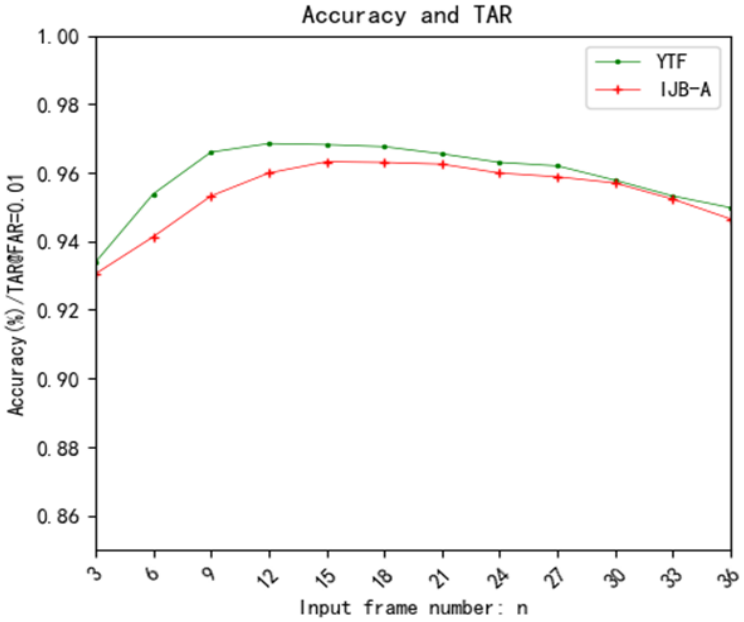

FIGURE 5. The influence of the number of input frames on the accuracy of the model

4.3. **Experimental evaluation based on dataset YTF.** Next, the proposed scheme is tested on the YTF dataset. The experimental results of the proposed method and the comparative model are shown in Table 2. Among them, VGGFace is mainly a recognition model for static face images, and its performance is poor when applied to video datasets, with an accuracy rate of only 92.41%. For the more advanced video face recognition models NAN and REAN, the recognition accuracy rates reached 95.72% and 96.60%, respectively. CNN+MaxPool has the worst recognition accuracy. In contrast, the method in this paper and CNN+AvePool increase the accuracy by 8.51% and 6.84%, respectively. REAN adopts the structure of recurrent neural network. QAN adds a branch of quality scores to the recognition network. FaceNet uses a deep convolutional neural network to learn to map images to Euclidean space, and spatial distance is related to image similarity. Compared with the frame structure-aware aggregation module in this paper, the recognition accuracy is not much different.

Table 2. Comparison of verification accuracy using YTF dataset

Methods	Accuracy (%)	Methods	Accuracy (%)
VGGFace	92.41	C-FAN [16]	96.50
REAN	96.60	QAN [17]	96.17
NAN	95.72	DAN	94.28
CNN+AvePool	95.18	Wen et al. [18]	94.90
CNN+MaxPool	88.34	FaceNet [19]	95.52
Ours	**96.85**	DeepID2+ [20]	93.20

It can be seen from Figure 5 that when the number of frames increases from 3 to 6, the recognition effect of the model in this paper is greatly improved; with the further increase of the number of frames, the effect of the model gradually improves, but it is not obvious. The best results are achieved when the number of input frames is 12, and the validation accuracy is 96.85%. When the number of frames exceeds 12 frames, the accuracy trend is almost unchanged, but still maintains a high recognition accuracy. It can be seen that since the method in this paper mainly focuses on the relationship and structural information between video frames, the number of input frames has a certain influence on it.

5. **Conclusions.** Aiming at the difficulty of face recognition in video, this paper proposes a frame structure-aware aggregation method combined with multi-scale feature extraction network from the perspective of decomposing video frames. The network effectively models the video context information, mines the position information contained in each frame of face image, and makes it affect the subsequent feature aggregation process. We conducted corresponding comparative experiments on the two public datasets, IJB-A and YTF, and proved that it has good performance. In the follow-up work, we will consider the segmentation strategy, divide the video frames into several ordered sets, and discuss the inter-relationships within and between sets to obtain more reasonable structure-aware information, thereby further improving the performance of the model.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (Nos. 62001133, 61661017, 61362021), Special Foundation for Scientific Bases and Talents of Guangxi Province (No. AD19110060), Guangxi One Thousand Young and Middle-aged College and University Backbone Teachers Cultivation Program, the National Natural Science Foundation of Guangxi Province, China (No. 2017GXNSF BA198212), the Fund of Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin University of Electronic Technology, (No. GXKL06200114), Innovation Project of GUET Graduate Education (2021YCXS027). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Wirianto and T. Mauritsius, The development of face recognition model in Indonesia pandemic context based on DCNN and Arcface loss function, *International Journal of Innovative Computing, Information and Control*, vol.17, no.5, pp.1513-1530, 2021.

[2] J. Wei and C. Ying, Aggregative adversarial network for still-to-video face recognition, *2020 5th International Conference on Computer and Communication Systems*, pp.266-270, 2020.

[3] M. Zhang, R. Liu, D. Deguchi and H. Murase, Context-aware contribution estimation for feature aggregation in video face recognition, *IEEE Access*, vol.10, pp.79301-79310, 2022.

[4] Y. Wang, Y. P. Huang and X. J. Shen, ST-VLAD: Video face recognition based on aggregated local spatial temporal descriptors, *IEEE Access*, vol.9, pp.31170-31178, 2021.

[5] J. Yang, P. Ren, D. Zhang et al., Neural aggregation network for video face recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.5216-5225, 2017.

[6] S. Gong, Y. Shi and A. K. Jain, Recurrent embedding aggregation network for video face recognition, *arXiv Preprint*, arXiv: 1904.12019, 2019.
[7] J. Rivero-Hernández, A. Morales-González, L. G. Denis et al., Ordered weighted aggregation networks for video face recognition, *Pattern Recognition Letters*, vol.146, no.2, pp.237-243, 2021.
[8] Q. Wang and G. Guo, LS-CNN: Characterizing local patches at multiple scales for face recognition, *IEEE Trans. Information Forensics and Security*, vol.15, pp.1640-1653, 2020.
[9] R. Hadsell, S. Chopra and Y. Lecun, Dimensionality reduction by learning an invariant mapping, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, pp.1735-1742, 2006.
[10] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv Preprint*, arXiv: 1412.6980, 2017.
[11] F. Brendan, B. Klein, E. Taborsky et al., Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1931-1939, 2015.
[12] L. Wolf, T. Hassner and I. Maoz, Face recognition in unconstrained videos with matched background similarity, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.529-534, 2011.
[13] K. Zhang, Z. Zhang, Z. Li et al., Joint face detection and alignment using multi-task cascaded convolutional networks, *IEEE Signal Processing Letters*, vol.23, no.10, pp.1499-1503, 2011.
[14] J.-C. Chen, V. M. Patel and R. Chellappa, Unconstrained face verification using deep CNN features, *IEEE Winter Conference on Applications of Computer Vision*, pp.1-9, 2016.
[15] O. M. Parkhi, A. Vedaldi and A. Zisserman, Deep face recognition, *British Machine Vision Conference*, pp.41.1-41.12, 2015.
[16] S. Gong, Y. Shi and A. K. Jain, Video face recognition: Component-wise feature aggregation network (C-FAN), *International Conference on Biometrics*, pp.1-8, 2019.
[17] Y. Lin, J. Yan and W. Ouyang, Quality aware network for set to set recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.4694-4703, 2017.
[18] Y. Wen, K. Zhang, Z. Li and Y. Qiao, A discriminative feature learning approach for deep face recognition, *European Conference on Computer Vision*, pp.499-515, 2016.
[19] F. Schroff, D. Kalenichenko and J. Philbin, FaceNet: A unified embedding for face recognition and clustering, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.815-823, 2015.
[20] Y. Sun, X. Wang and X. Tang, Deeply learned face representations are sparse, selective, and robust, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2892-2900, 2015.