A DUAL-MODALITY HUMAN FEATURE RECOGNITION METHOD USING MATCHED LAYER FUSION

Ke DENG [1], Jiawei MO[*] [2], Rongping HUANG [3]

The objective of this paper is to present a recognition method based on the fusion of two human features, namely gait and face, within a matching layer. Space-temporal biometric features with differentiation in human contour maps are obtained by gait feature extraction network in order to resolve the issue that facial recognition technology is difficult to recognize the target subject with high accuracy under the condition of having interfering objects on the face or longer detection distance. A facial feature extraction network is employed to obtain fine-grained features of the face in order to enhance the immunity of the network to the conditions where the contour of the target subject is affected by interfering objects. Facial features are fused with gait features at the matching layer for information fusion in order to achieve complementarity between the two modal biometrics. The experimental results show that the method proposed in the paper has higher recognition accuracy compared to the gait or facial feature recognition methods in unimodal mode.

Keywords: modality; feature fusion; facial recognition; gait recognition

1. Introduction

Facial recognition technology has become an important research topic [1] in the field of identification in recent years. Facial recognition is a common identification technology that is employed in a variety of contexts, including access control, online payment and system login, and numerous other fields [2]. The using of facial recognition techniques for the purpose of identification offers a number of advantages, including the ability to perform the identification process without direct contact, in a covert manner, at a high rate of speed, and with a high degree of efficiency. Nevertheless, the efficacy of facial recognition is constrained by the actual distance between the camera and the face. In numerous recognition application scenarios, it is not feasible to recognize the target subject at a greater distance. Furthermore, when the face is obstructed by an occluding object (e.g., when the target subject is wearing a mask or the brim of a hat is positioned too low, blocking the eyebrows, etc.), the accuracy of recognition is likely to be lower than that required by the application.

Gait is a biological trait that describes the manner in which an individual walk. Gait features can be observed at greater distances than facial features and do

[1] Senior Engineer, Liuzhou Institute of Technology, Liuzhou, China;
[2] Senior Engineer, Liuzhou Institute of Technology, Liuzhou, China, *Corresponding author's e-mail: 23624612@qq.com;
[3] Lecturer, Liuzhou Institute of Technology, Liuzhou, China.

not require the cooperation of the target [3]. Currently, the majority of gait recognition methods use contour-based gait recognition networks [4]. The target's silhouette is susceptible to distortion due to the presence of external factors such as objects being carried or loose clothing. When the contour of the target subject is obscured by interfering with objects, the accuracy of gait recognition is significantly reduced.

A recognition method based on the fusion of two modal features, gait and face, in a matching layer is proposed in this paper. The feature fusion through the matching layer achieves the effect of complementary information. For the extraction of the gait feature, a global-local space-temporal feature extraction module is constructed. Furthermore, the local feature extraction module employs a fine-grained feature extraction strategy and a complementary mask-based multi-scale random band segmentation approach to enhance the correlation between each local feature. The global information extracted by the global feature extraction module is employed to enhance the resilience of the gait feature extraction network.

2. Related work

2.1. Facial features recognition

The advancement of deep learning theory has led to a significant enhancement in the accuracy of facial recognition technology, rendering it one of the most prevalent forms of biometrics. Chan et al [5]. proposed a deep learning network, PCA-Net, with a simple structure. This network employs a PCA filter to filter the features of interfering objects, thereby reducing the influence of these objects on facial recognition. Schroff et al [6]. proposed Face-Net, which introduces a ternary loss function. This allows the extracted features to demonstrate the property of smaller intra-class distances and larger class-to-class spacing. This makes facial features easier to distinguish. Li et al [7]. introduced adversarial generative networks into facial recognition for repairing the disturbed region, but the effect of repairing the detailed features of the face is average. Amos et al [8]. constructed the Open-face code base for facial recognition, which greatly facilitated the development of facial recognition. Practice has shown that unimodal-based facial recognition techniques are prone to inaccuracy when attempting to identify targets that are distant or obscured by masks.

2.2. Gait features recognition

There are two broad categories of gait recognition techniques, depending on the form of the input data: those based on skeleton features and those based on human contour features. The majority of skeletal feature-based recognition methods [9-12] are based on the human pose estimation algorithm [13], which takes the detected human skeletal node information as input data. Although this method is highly resistant to interfering objects, there are certain errors in the skeleton data

extraction process, resulting in lower recognition accuracy than the human contour-based recognition method. Human silhouette feature-based recognition methods use the human silhouette set as input data to learn useful space-temporal representations of gait. The form of characterization of gait features can be further divided into global feature characterization and local feature characterization. For example, Chao et al [14]. proposed Gait-Set, which treats the entire human body as a single entity and employs 2D convolution to extract feature representations of contour images. In their study, Lin et al [15]. employed 3D convolution for the extraction of space-temporal features from contour images in gait sequences. Although the global-based feature characterization method has the advantages of low computation and high accuracy, the method can easily ignore the local fine-grained features of human posture as the network becomes deeper and deeper. Fan et al [16]. proposed Gait-Part to extract a more fine-grained gait representation by horizontally segmenting the contours in the sequence set. Zhang et al [17]. divided the entire gait profile into four localities and subsequently used 2D convolution to extract fine-grained local features. Although the method based on local gait characterization can effectively capture local fine-grained gait information, it does not learn the correlation between different local features and also requires a predefined segmentation strategy for a specific dataset, which results in a model that is less effective and versatile for facial recognition when there are interfering objects influencing it.

2.3. Multi-modal feature recognition

The existing human body recognition based on multimodal fusion is generally a variety of human features that are fused at different levels and then recognized. Depending on the level of fusion, this can be divided into four main categories: fusion at the data level, fusion at the feature level, fusion at the matching level and fusion at the decision level [18]. The primary objective of multimodal fusion is to diminish the disparate characteristics between modalities, thereby facilitating complementarity, while maintaining the distinctiveness of each modality's semantics and enhancing the efficacy of multimodal human body recognition [19]. Soltani et al [20]. spliced and fused multiple fingerprint images at the data layer to achieve fingerprint fusion recognition at the data layer. Soleymani et al [21]. employed an enhanced Fisher classifier for serial feature fusion. The research of Wang et al [22]. proposed a feature layer fusion strategy that is based on typical correlation analysis. Muthukumaran et al [23]. used a classifier-based decision fusion approach, where the matching distances of the face and iris classifiers are considered as a two-dimensional feature vector, which is then classified as true or false using classifiers such as Fisher's Discriminant Analysis or Neural Networks with Radial Basis Functions (RBFNN). Mustafa et al [24]. proposed a transform-based matching layer fusion method for synthetics the final

fused scores after normalization the different modal scores to the same interval. Given that gait is a human feature represented in the form of a video, and that face is a human feature represented in the form of an image, the fusion in the pixel layer as well as in the feature layer is subject to incompatibility of features, which in turn makes it difficult for the network to be trained and learnt effectively. The matching layer fusion can better balance the difficulty of raw information and data processing, maximum the complementary fusion of human features from different modalities and improve the stability and training efficiency of the network. Therefore, this paper employs a transform-based matching layer fusion method to integrate the gait and facial feature data.

3. Methods

The principle flow of the dual-modality human feature recognition method proposed in the paper is shown in Figure 1. The model comprises two distinct branches: a gait feature extraction network and a facial feature extraction network.

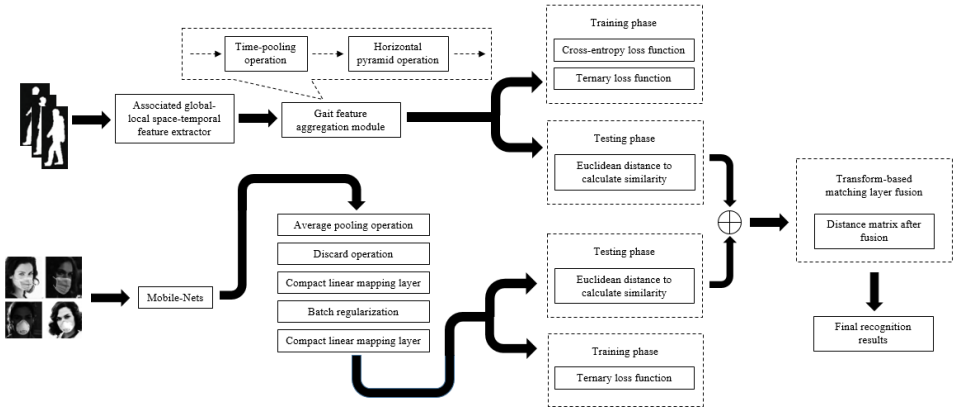

Fig. 1. Procedure of the proposed method.

The gait feature extraction network is employed to achieve accurate long-range recognition of the target subject and to enhance the gait single branch's immunity to occlusions to a certain extent via a multi-scale random band segmentation strategy based on complementary masks. The facial feature extraction network is used to extract fine-grained facial features to complement the gait features. This integration enhances the model's resilience to interference when the gait profile is obstructed. The distance matrix, derived from gait and facial branching, is normalization at the matching layer, after which information fusion is performed. Finally, based on the fused distance matrix, Rank-1 evaluation metrics are used to obtain the final recognition results.

3.1. Facial feature extraction network

The paper uses the Face-Net [5] method to construct this network and Mobile-Nets [25] as the backbone network to extract facial features. The network used to extract facial features consists of a batch input layer and a Mobile-Nets backbone network. Next, a compact linear mapping layer is used to obtain the facial features and a ternary loss function is used for training. The input features $I \in R^{c_1 \times h_1 \times w_1}$ are obtained through a batch input layer, where $h \times w$ denotes the size of each frame. The Mobile-Nets backbone network was then used to extract distinguishable facial features $F \in R^{c_2 \times h_2 \times w_2}$. Finally, a compact linear mapping layer is used to extract the D-dimensional facial features F_{face} ,the calculation process is shown in equation (1).

$$F_{face} = L_{compactlayer}(MobileNets(I)) \tag{1}$$

3.2. Gait feature extraction network

The 3D convolutional block simultaneously extracts space-temporal features of the gait sequence, thus simplifying the previous process of extracting temporal and spatial features separately. Furthermore, it preserves the complete gait periodicity features to a certain extent. LIN et al. proposed Gait-GL [26], a method that employs 3D convolution for the extraction of both global and local features. This approach involves the segmentation of the input features horizontally, allowing for the extraction of fine-grained features within each local region. HUANG et al. proposed 3D Local [27], which extracts the limb features of the target subject through adaptive scale 3D local convolution operation. Nevertheless, none of these methods is capable of effectively capturing the correlation information between neighboring local regions, which consequently limits the representation of local gait features. This paper proposes the corresponding global-local space-temporal feature extractor, which has the structure of Fig. 2. The module employs a multi-scale random band segmentation strategy based on complementary masks and using 3D convolution to extract global gait space-temporal features. This approach enables the effective extraction of correlation information between neighbor local features, the full excavation of global-local space-temporal features with differentiation and improve the model's immunity to occlusions. Finally, the gait feature aggregation module performs a pooling operation on the sequence information in the time dimension, resulting in the generation of the final gait features through the linear mapping layer. The input to the gait feature extraction network is assumed to be a series of contour sequences $I \in R^{1 \times f \times h \times w}$. Where the number of channels in the gait profile is 1 and f is the number of frames in the sequence. The associated global-local space-temporal feature extractor can be expressed in equation (2).

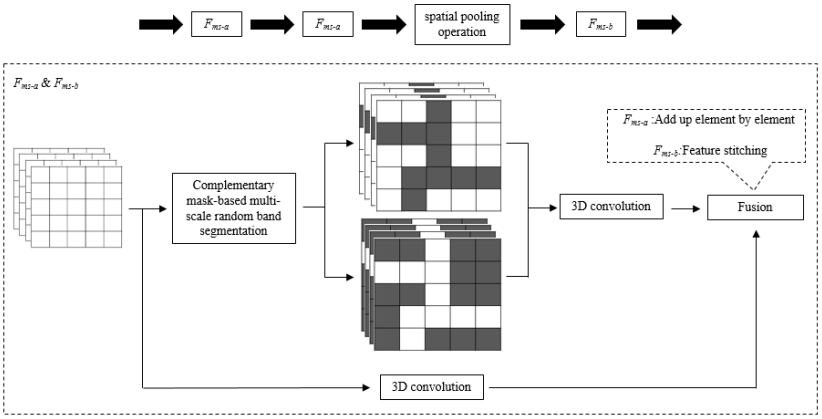

Fig. 2. Flow chart for recognizing indication value of pointer meters.

$$F_{rgl} = F_{ms-b}(F_{sp}(F_{ms-a}(F_{ms-a}(I)))) \tag{2}$$

Where: $F_{rgl} \in R^{c_1 \times t \times \frac{1}{2}(h \times w)}$; F_{sp} denotes the spatial pooling operation, subsampled the features, keeps the channel and frame values constant, and reduces the product of height and width by 1/2. $F_{ms\text{-}a}$ and $F_{ms\text{-}b}$ denote the global-local feature extraction module for complementary mask-based multiscale stochastic band segmentation, where the former fuses global and local features by direct summation, and the latter splices global and local features in H-dimension to obtain fused features. The calculation process for $F_{ms\text{-}a}$ and $F_{ms\text{-}b}$ can be expressed as shown in equations (3) and (4).

$$F_{ms-a} = F_{global}(I) + F_{local}(I) \tag{3}$$

$$F_{ms-b} = Concat(F_{global}(I), F_{local}(I)) \tag{4}$$

Where: F_{global} denotes the 3D convolution operation; F_{local} denotes the more detailed representation of gait information in feature F_m and its complementary feature $\overline{F_m}$, which have been extracted using weight-sharing 3D convolution after the random band mask has been applied. At the same time, the presence of complementary features enables the module to effectively extract the correlation information between neighboring local features and improve the network's immunity to occlusions. The equations from (5) to (8) represent the computational processes of F_{global}, F_m, $\overline{F_m}$ and F_{local}, respectively.

$$F_{global} = 3DConv(I) \tag{5}$$

$$F_m = W_{mask} \otimes I \tag{6}$$

$$\overline{F_m} = (1 - W_{mask}) \otimes I \tag{7}$$

$$F_{local} = 3DConv(F_m) \oplus 3DConv(\overline{F_m}) \tag{8}$$

The multi-scale random band mask W_{mask} and its complementary mask structure are shown in Fig. 3. In the height and width dimensions, the paper introduces an approach that enriches the diversity of input features by applying

multi-scale random band masking to the feature maps in order to simulate a real scenario where important local features of the human body are occluded. Concurrently, the presence of complementary features ensures that crucial gait feature information is not lost, thereby facilitating the model proposed in the paper to discern correlations between local features.

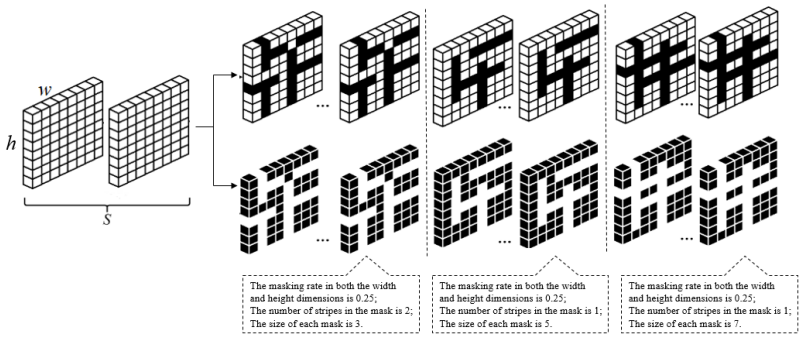

Fig. 3. Complementary mask-based multi-scale random band segmentation approach.

The gait feature aggregation module consists of a time pooling layer as well as a horizontal pyramid mapping layer proposed in the references [14], and the computational process is shown in Equation (9).

$$F_{gait} = H_{hpm}(T_{tp}(F_{rgl})) \tag{9}$$

Where: T_{tp} denotes time pooling operation; H_{hpm} denotes the horizontal pyramid mapping operation; F_{gait} denotes the final gait feature data.

3.3 Matching layer fusion method

Feature fusion based on pixel or feature layers can result in incompatibility between the gait features and facial features, which in turn presents a challenge for the network in terms of training and learning effectively. The paper opts to conduct a complementary fusion of gait feature information with facial feature information at the matching layer. Given that the paper employs a dual-modality distance information fusion at the matching layer, it is only necessary to train the modal backbone network individually. Furthermore, the parallelize of model training can be achieved without decoupling each modal network simultaneously. Subsequently, following the convergence of each branch network, dual-modality fusion retrieval is conducted. The dual-modality human feature fusion retrieval mechanism, based on a matching layer, is illustrated in Fig. 4.

During the testing phase, facial images are input into a trained facial recognition network in order to extract the facial features F_{face}. The gait profile is input into the gait feature extraction network, as outlined in the paper, in order to obtain the gait feature F_{gait}. The paper employs the Euclidean distance, as

illustrated in Equation 10, to quantify the similarity between the individual to be verified V_i, and the candidate C_j, in the facial and gait test sets.

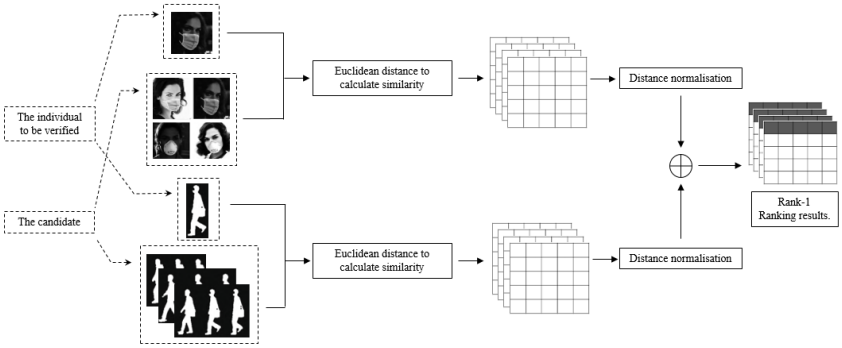

Fig. 4. Dual-modality human feature recognition structure using matched layer fusion.

$$d(V_i, C_j) = \sqrt{(V_i - C_j)^2} \tag{10}$$

where: the person to be verified V_i denotes the individual waiting for an identity test and the subscript i is a positive integer; Candidate C_j denotes an individual with the same label as the person to be verified V_i in the comparison database, and the subscript j is a positive integer. Given that the distance magnitude between the individual to be verified and the candidate varies depending on the modality in question, it is necessary to normalization the distances in order to ensure that the impact of different modalities on the decision is balanced. This process is illustrated in Equation 11.

$$d\hat{}(V_i, C_k) = \frac{d(V_i, C_k)}{\sum_{j=1}^{n} d(V_i, C_j)} \tag{11}$$

The transform-based matching layer fusion method normalization the gait similarity distance and face similarity distance in order to obtain D_{face} and D_{gait}, respectively. Ultimately, as illustrated in Equation 12, the face and gait similarities are aggregated to derive the fusion metric D_{fuse}.

$$D_{fuse} = D_{face} + D_{gait} \tag{12}$$

4. Experiments and results

The proposed dual-modality human feature recognition methods are implemented using the Python and PyTorch deep learning network frameworks.

4.1. Selected dataset

Datasets CASIA-WebFace [28], CASIA-B [29] and CASIA-WebMaskedFace [30] were selected for analysis in this paper.

CASIA-WebFace: This dataset is a common resource for facial recognition model training, comprising 494,414 facial images from 10,575 individuals. In order to facilitate the subsequent construction of the dual-modality dataset, the thesis will

select the subset consisting of the first 10,075 people in this dataset as the training set for the facial recognition model, and the subset consisting of the next 500 people as the test set for the facial recognition model.

CASIA-B: This dataset is currently the most widely used gait dataset, comprising 124 subjects. Each subject comprised 11 angles, with each angle containing 10 sequences. Three walking states were identified for each sequence: the normal state (*N*), the state of carrying items (*B*), and the state of wearing loose clothing (*C*). In total, the dataset comprises 13,640(124×11×10) videos. The paper used contour sequences as gait input data, with the first 74 subjects as the training set and the next 50 subjects as the test set. The first 4 sequences in the normal state are kept in the candidate set as candidates (i.e., N1~N4) and the remaining 6 sequences are kept in the set of persons to be verified (i.e., N5, N6, B1, B2, C1, C2). The dataset is employed as the training set for the gait recognition model and the test set for the dual-modality fusion model. The sequence of gait profiles is illustrated in Fig. 5.

(a) the normal state (*N* conditions)

(b) the state of carrying items (*B* conditions)

(c) the state of carrying items (*C* conditions)

Fig. 5. Sequence of gait profiles for each state.

CASIA-WebMaskedFace: The dataset is based on CASIA-WebFace, and a mask masking effect was added to the facial images in the dataset using the MaskTheFace tool. This was done in order to mimic distractors. The effect is shown in Fig. 6.

(a) medical surgical masks

(b) medical surgical masks

(c) N95 masks

(d) N95 masks

Fig. 6. Effectiveness of different types of masks in covering the face.

The paper employed a random selection of mask types (e.g., medical surgical masks, N95 masks, etc.) to simulate the impact of masking. The final 500 individuals from the CASIA-WebMaskedFace dataset were employed as the test dataset for the facial recognition model in the mask-wearing condition.

For the dual-modality fusion experiments, the dual-modality test dataset (test samples containing both gait and facial feature data) constructed in the paper was divided into two categories in order to correspond to the size of the CASIA-B gait dataset. In the first category, 50 subjects from the 500 test samples in CASIA-WebFace were selected to form the face-unmasked dual-modality test dataset with the 50 test samples in CASIA-B. In the second category, 50 subjects were randomly selected from CASIA-WebMaskedFace and combined with the test set in CASIA-B to form the face-masked of the dual-modality test dataset. Table 1 lists the statistics of the number of samples in each type of dataset. The dual-modality dataset constituting the thesis contains occluded facial pictures as well as gait contours.

Table 1

Statistics on the number of datasets.

datasets	training set	test set
CASIA-WebFace	10075	500
CASIA-WebMaskedFace	-	500
CASIA-B	74	50
CASIA-B+ WebFace	-	50
CASIA-B+ WebMaskedFace	-	50

4.2. Experimental parameter settings

The facial feature extraction network uses Mobile-Nets network as a backbone network to extract facial features. During the training process, the paper uses the momentum-based Adam optimization process to train the entire end-to-end facial recognition network, with momentum set to 0.9, a cosine annealing learning rate strategy with a maximum learning rate of 1×10^{-3}, a minimum learning rate of 1×10^{-5}, a total number of epochs of 100, and a batch size (Batch Size) of 96. The paper uses the Rank-1 evaluation metric for facial recognition accuracy in order to be consistent with the evaluation criteria in the field of gait recognition when testing the facial feature extraction network.

The gait feature extraction network uses the Adam optimization process with weight decay set to 5×10^{-4}. The MultiStepLR learning rate strategy is used with an initial learning rate of 1×10^{-4}, a learning rate set to 1×10^{-5} after 7×10^{4} iterations, a total number of iterations of 8×10^{4}, and a batch size (Batch Size) of 8×16, where m in the ternary loss is set to 0.2.

4.3. Results

Table 2 presents a comparison of the recognition accuracy of various gait feature extraction networks in the CASIA-B dataset. In the experiment of gait feature extraction networks, the network based on a correlated global-local space--temporal feature extractor, as proposed in the paper, achieved 92.6% recognition accuracy in the CASIA-B dataset. Compared with the baseline network Gait-GL[26], it improves by 0.8 percentage points, especially by 1.6 percentage points when the

person to be verified is wearing loose clothing, indicating that the introduction of the complementary mask-based multi-scale random band segmentation strategy can sufficiently learn the correlation information between different local regions, and improve the network's immunity to interference for occluding objects to a certain extent, so that the model can be applied to more complex scenarios.

Table 2

Recognition accuracy of various gait feature extraction networks in the CASIA-B dataset

datasets	State *N* (%)	State *B* (%)	State *C* (%)	Average values (%)
Gait-Set	95.0	87.2	70.4	84.2
Gait-Part	96.2	91.5	78.7	88.8
Gait-GL	97.4	94.5	83.6	91.8
3D-Local [27]	**97.5**	94.3	83.7	91.8
Methods of this paper	**97.5**	**95.2**	**85.2**	92.6

Table 3 illustrates the comparison of the recognition accuracy of facial feature extraction networks on the CASIA-WebFace and CASIA-WebMaskedFace [30] datasets under different conditions.

Table 3

Recognition accuracy of facial feature extraction networks under different conditions

Resolution of facial images (pixels)	CASIA-WebFace (%)	CASIA-WebMaskedFace (%)
72 × 72	92.34	88.54
96 × 96	96.21	92.22
112 × 112	98.71	95.67

In the CASIA-WebFace dataset, the facial feature extraction network described in the paper achieves 98.71% accuracy in recognizing facial images with a resolution of 112×112 pixels. Nevertheless, as the resolution of the facial image decreases (the recognition distance increases), the accuracy of the recognition also decreases. A reduction in the resolution of the facial image to 72×72 pixels resulted in a decline in recognition accuracy to 92.34%, a decrease of 6.37 percentage points in comparison to the facial image resolution of 112×112 pixels. Compared to the unobscured face dataset CASIA-WebFace, the recognition accuracy of the facial feature extraction network in the CASIA-WebMaskedFace dataset with facial occlusion decreases by 3.04, 3.99, and 3.80 percentage points when the resolution of the facial image is 112×112 pixels, 96×96 pixels, and 72×72 pixels, respectively. The results demonstrate that the unimodal facial feature extraction network exhibits a notable decline in the accuracy of target subject recognition when the face is occluded. In light of the aforementioned limitations, this paper proposes a dual-modality human feature recognition method. This approach is designed to address the issue of low recognition accuracy associated with unimodal facial feature extraction networks.

Table 4 illustrates the recognition accuracy of the dual-modality human feature recognition network under different conditions. The experimental results obtained from CASIA-B+CASIA-WebFace demonstrate that the accuracy of the

dual-modality human feature recognition network has been enhanced from 92.6% to 99.16%. This represents an improvement of 15% compared to the original unimodal gait feature extraction network (The first row of data in Table 2 of this paper). Furthermore, the method demonstrates enhanced robustness with regard to recognition accuracy under *C* conditions. This approach mitigates the sensitivity of unimodal gait feature extraction networks to interfering objects that occlude the limbs, thus exploiting the complementary strengths of facial and gait features. From the experimental results of CASIA-B+CASIA-WebMaskedFace, it can be seen that the dual-modality human feature recognition network is still able to achieve 94.52% accuracy even at a long distance (when the facial image resolution is 72×72 pixels) and when the face is covered by a mask, which is an improvement of 5.98 percentage points compared to the accuracy of the facial feature extraction network on the CASIA-WebMaskedFace dataset.

Table 4

Recognition accuracy of the dual-modality human feature recognition network under different conditions

Datasets	Resolution of facial images (pixels)	State *N* (%)	State *B* (%)	State *C* (%)	Average values (%)
CASIA-B+ WebFace	72 × 72	98.38	96.81	90.89	95.36
	96 × 96	98.96	98.15	93.92	97.01
	112 × 112	99.62	99.30	98.55	99.16
CASIA-B+ WebMaskedFace	72 × 72	98.12	96.44	89.00	94.52
	96 × 96	98.48	97.06	89.71	95.08
	112 × 112	98.44	96.99	91.42	95.62

The bimodal human feature recognition method proposed in the paper, which employs a facial image resolution of only 72×72 pixels and no facial occlusion, exhibits an accuracy of 95.36%. This value represents a 3.04 percentage point improvement over the accuracy of the facial feature extraction network under identical conditions. The recognition accuracy of the proposed method in the paper is improved by 5.98 percentage points compared to the facial feature extraction network in the case of occluded faces and a facial image resolution of 72×72 pixels. The experimental results demonstrate that the incorporation of a gait feature extraction network into the model enables the effective recognition of a target wearing a mask over a long distance, effectively addressing the issue of the significant decline in accuracy of the unimodal recognition method in complex application scenarios.

5. Conclusion

The paper puts forth a dual-modality human feature recognition method based on matching layer fusion. By fusing gait features with facial features, it solves the problem that unimodal facial recognition methods in real scenarios are difficult

to accurately recognize targets with occluding objects on the face or at long distances. Concurrently, the resilience of the gait feature extraction network to alterations in appearance resulting from different wearing styles is enhanced by integrating facial and gait features in a complementary manner. In order to enhance the adaptability of the gait profile to occlusion and viewpoint changes, this paper proposes to design an associated global-local space-temporal feature extraction module for the gait feature extraction network. The experimental results show that the method is not only able to perform information fusion with high accuracy under typical conditions but is also able to accurately identify the target in the presence of occluding objects on the face or in distant scenes.

Acknowledgment

This work was supported in part by the project on enhancement of basic research ability of young and middle-aged teachers in Guangxi universities and colleges (2023KY1774), in part by the Guangxi higher education undergraduate teaching reform project (2023JGZ187) in China.

R E F E R E N C E S

[1]. *M. Gomez-Barrero, P. Drozdowski, C. Rathgeb*, "Biometrics in the Era of COVID-19: Challenges and Opportunities" in IEEE Transactions on Technology and Society, vol. 3, no. 4, pp. 307-322, Dec. 2022.

[2]. *D. Osorio-Roig, C. Rathgeb, P. Drozdowski, P. Terhörst, V. Štruc*, et al., "An Attack on Facial Soft-Biometric Privacy Enhancement" in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 4, no. 2, pp. 263-275, April 2022.

[3]. *S. U. Yunas, K. B. Ozanyan*, "Gait Activity Classification from Feature-Level Sensor Fusion of Multi-Modality Systems" in IEEE Sensors Journal, vol. 21, no. 4, pp. 4801-4810, 15 Feb.15, 2021.

[4]. *R. Ryu, S. Yeom, S. -H. Kim, D. Herbert*, "Continuous Multimodal Biometric Authentication Schemes: A Systematic Review" in IEEE Access, vol. 9, pp. 34541-34557, 2021.

[5]. *T.-H. Chan, K. Jia, S. Gao*, "PCANet: A Simple Deep Learning Baseline for Image Classification?" in IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5017-5032, Dec. 2015.

[6]. *F. Schroff, D. Kalenichenko, J. Philbin*, "FaceNet: A unified embedding for face recognition and clustering" in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 815-823.

[7]. *Yijun Li, Sifei Liu, Jimei Yang, Ming-Hsuan Yang*, "Generative Face Completion" in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 5892-5900.

[8]. *B. Amos, B. Ludwiczuk*, "OpenFace: A general-purpose face recognition library with mobile applications" (2016). CMU-CS-16-118, CMU School of Computer Science.

[9]. *LI X, Yasushi Makihara, Chi Xu, Yasushi Yagi*, "End-to-end model-based gait recognition", Proceedings of the 15th Asian Conference on Computer Vision. Berlin, Germany: Springer, 2020:1-10.

[10]. *Junjie Huang, Zheng Zhu, Guan Huang*, "Multi-Stage HRNet: Multiple Stage High-Resolution Network for Human Pose Estimation". ArXiv abs/1910.05901 (2019): n. pag.

[11]. *T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann*, "Gaitgraph: graph convolutional network for skeleton-based gait recognition" Proceedings of International Conference on Image Processing. Washington D.C., USA: IEEE Press, 2021:2314-2318.

[12]. *Rijun Liao, Shiqi Yu, Weizhi An*, "A model-based gait recognition method with body pose and human prior knowledge" in Pattern Recognition, 2020, 98:107069.
[13]. *Xiaokai Liu, Zhaoyang You, Yuxiang He*, "Symmetry-Driven hyper feature GCN for skeleton-based gait recognition" in Pattern Recognit. 125 (2022): 108520.
[14]. *H.J. Chao, Y.W. He, J.P. Zhang*, "GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition" in Proceedings of the AAAI Conference on Artificial Intelligence, 2019,33(01), 8126-8133.
[15]. *B. Lin, S. Zhang, Y. Liu and S. Qin*, "Multi-Scale Temporal Information Extractor for Gait Recognition" in 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021, pp. 2998-3002.
[16]. *C. Fan,Y. Peng, C. Cao,X. Liu*, "GaitPart: Temporal Part-Based Model for Gait Recognition" in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 14213-14221.
[17]. *Y. Zhang, Y. Huang, S. Yu, L. Wang*, "Cross-View Gait Recognition by Discriminative Feature Learning" in IEEE Transactions on Image Processing, vol. 29, pp. 1001-1015, 2020.
[18]. *Z.Y. GAO, B. LI*, "Face and iris fusion recognition based on Adaboost" in Computer Engineering, 2011, 37(6):148-150.
[19]. *J. Guo, Q. Liu and E. Chen*, "A Deep Reinforcement Learning Method for Multimodal Data Fusion in Action Recognition" in IEEE Signal Processing Letters, vol. 29, pp. 120-124, 2022.
[20]. *R. Soltani, D. Goeckel, D. Towsley*, "Fundamental Limits of Invisible Flow Fingerprinting" in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 345-360, 2020.
[21]. *S. Soleymani, A. Dabouei, F. Taherkhani, S. M. Iranmanesh*, "Quality-Aware Multimodal Biometric Recognition" in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 4, no. 1, pp. 97-116, Jan. 2022.
[22]. *M. Wang, Y. Liu, W.F. Liu & B.D. Liu*, "Feature Fusion Based Parallel Graph Convolutional Neural Network for Image Annotation" in Neural Processing Letters 55 (2023): 6153-6164.
[23]. *B. Muthukumaran*, "Face and Iris based Human Authentication using Deep Learning" in 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2023, pp. 841-846.
[24]. *A. S. Mustafa, A. J. Abdulelah*. "Multimodal Biometric System Iris and Fingerprint Recognition Based on Fusion Technique" in International journal of advanced science and technology. Vol 29, No.23 (2020), pp7423-7432.
[25]. *Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". ArXiv abs/1704.04861 (2017): n. pag.
[26]. *B. Lin, S. Zhang, X. Yu*, "Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation" in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 14628-14636.
[27]. Z. Huang, D. Xue, X. Shen, X. Tian, "3D Local Convolutional Neural Networks for Gait Recognition" in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 14900-14909.
[28]. *Y. Dong, L. Zhen, S.C. Liao*, "Learning Face Representation from Scratch". ArXiv abs/1411.7923 (2014): n. pag.
[29]. *A. Bansal, A. Jain, S. Bharadwaj*, "An Exploration of Gait Datasets and Their Implications" in 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2024, pp. 1-6.
[30]. *T. A. Mare, G. Duta*, "A realistic approach to generate masked faces applied on two novel masked face recognition data sets" in 35th Conference on Neural Information Processing Systems, Oct. 2021.