Dynamic Hand Gesture Authentication Based on Improved Two-Stream CNN

Wenwei Song[1,2], Linpu Fang[1], Yihong Lin[1,2], Ming Zeng[1], and Wenxiong Kang[1,2,3(✉)]

[1] School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China
auwxkang@scut.edu.cn
[2] Pazhou Lab, Guangzhou 510335, China
[3] Guangdong Enterprise Key Laboratory of Intelligent Finance, Guangzhou 510705, China

Abstract. Recently, dynamic hand gesture (DHG) has been discovered to be a promising biometric trait containing both physiological and behavioral characteristics simultaneously. DHGs are recorded by videos, so the authentication process is a challenging fine-grained video understanding task. Fully exploring physiological and behavioral characteristics to capture the fine-grained spatiotemporal identity features is the key to DHG authentication. Thus, in this paper, we propose to use the classic two-stream CNN for video understanding-based DHG authentication due to their explicit spatial (static) and temporal (dynamic) information modeling ability. Through analyzing prior two-stream CNN-based authentication methods in depth, we find five improvable aspects and propose the corresponding enhancement strategies. Comprehensive experiments on the SCUT-DHGA dataset show that the improved two-stream CNN can significantly outperform existing SOTA DHG authentication methods.

Keywords: Biometrics · Dynamic hand gesture authentication · Two-stream CNN · Video understanding

1 Introduction

With the great development of intelligent devices, it is urgent to develop easy-to-use and secure user authentication methods for access control. Traditional knowledge-based authentication methods such as passwords cannot meet the requirement of convenience and security due to the increases of memory burden and the risks of malicious attacks [1]. In recent years, biometric authentication methods have been proposed as effective alternatives to traditional knowledge-based counterparts and have been extensively used in various products such as intelligent phones and laptops. Biometric authentication methods utilize distinctive biometric traits to verify users' identities. Since the biometric traits are unique and unforgettable, they can bring users a more secure and user-friendly experience.

Among existing biometric traits, DHGs are very promising. Compared with physiological characteristic dominated traits, such as faces, fingerprints, irises, finger veins, *etc.*, DHGs can provide stronger resistance to spoofing attacks with the help of behavioral characteristics. Moreover, DHGs also contain abundant physiological

W. Deng et al. (Eds.): CCBR 2022, LNCS 13628, pp. 237–246, 2022.
https://doi.org/10.1007/978-3-031-20233-9_24

characteristics, such as hand shape and skin textures. The physiological and behavioral characteristics are highly complementary, thus making DHG authentication more accurate and more secure.

DHG authentication systems mainly use three kinds of acquisition devices to capture hand gestures currently, including inertial measurement units [2,3], touch screens [4,5], and cameras [1,6,7]. The inertial measurement unit and the touch screen require users to contact the devices physically, so the intrusive data collection manners significantly reduce the user-friendliness of DHG authentication. Also, the captured hand gestures are usually represented as trajectories of fingers' movements. Thus, a large number of physiological characteristics are lost, which is not conducive to taking full advantage of the DHGs. The video-based DHG authentication system uses a camera to obtain gesture videos. This non-contact way is more user-friendly. Besides, DHG videos retain plentiful physiological and behavioral characteristics. Therefore, video-based DHG authentication is more desirable.

Video-based DHG authentication methods are mainly divided into two categories: trajectory analysis-based methods and video understanding-based methods. The trajectory analysis-based methods first estimate the motion trajectories of hand keypoints [1,6,8,9] from videos with hand pose estimation algorithms [10] and then extract behavioral features from these motion trajectories. The video understanding-based methods directly extract features from videos that contains sufficient physiological and behavioral characteristics. The physiological characteristics are embedded in each frame, while behavioral characteristics are embodied in successive adjacent frames. Considering that video understanding based-methods can obtain richer identity features, in this paper, we focus on the video understanding-based DHG authentication.

For video understanding-based DHG authentication, the key is to distill fine-grained spatiotemporal identity features. Wu *et al.* [7] first extracted the silhouette covariance descriptor from DHG videos to verify users; however, the performance of this hand-crafted feature extraction method is not satisfactory. Wu *et al.* [11] afterward adopted a powerful two-stream convolutional neural network (TS-CNN) [12]) to extract features, resulting in significant performance improvement. Recently, Liu *et al.* [13] released a large-scale DHG authentication dataset called SCUT-DHGA, and proposed the DHGA-net based on I3D (a 3D CNN) [14], to benchmark this dataset. The results demonstrate that the DHGA-net is significantly superior to the TS-CNN in terms of EER (Error Equal Rate). Whereas, the interpretability of 3D CNNs is worse than two-stream CNNs since they jointly learn the physiological and behavioral characteristics, hardly figuring out the effects of the two parts.

In this paper, to fully unleash the potential of two-stream CNNs in DHG authentication, we analyze them in depth and then adopt five strategies to improve them. The results on the SCUT-DHGA dataset indicate that our enhanced two-stream CNN can not only achieve significant performance improvement compared with the TS-CNN but also can outperform the DHGA-net by a large margin.

2 The Proposed Method

2.1 Analyses of the Two Stream CNN-Based DHG Authentication Method

Through careful analyses of the two stream CNN-based DHG authentication method (TS-CNN) [11], we think it mainly has five improvable drawbacks. (1) **Hand gesture videos are underutilized:** The TS-CNN uses the single RGB image/depth map and single optical flow image broken up from videos as training samples to train the spatial and temporal branches, respectively, which greatly increases the difficulty of network learning. (2) **An inefficient optical flow is used:** The extraction of optical flow is very time-consuming and storage-demanding. Moreover, the optical flow is often pre-calculated offline, which cannot achieve adaptive behavior/motion representation learning according to specific tasks. (3) **An obsolete backbone is adopted:** The TS-CNN uses the classical AlexNet as the backbone for image feature extraction. However, the AlexNet not only has a larger number of model parameters but also performs worse than current advanced network architectures, such as the ResNet. (4) **A suboptimal loss function is selected:** DHG authentication can be regarded as a metric learning task. The TS-CNN uses the Softmax loss as the supervision signal to train the model, which has been proved that it is less effective in some metric learning tasks, such as face authentication [15]. (5) **The information fusion strategy is not fully explored:** Reasonable fusion of physiological and behavioral information is significant for improving hand gesture authentication performance. The TS-CNN only uses the intermediate-fusion strategy, which may not be the best choice currently, as it is challenging to achieve satisfactory feature fusion.

2.2 Improved Two-Stream CNN

For the first drawback of the TS-CNN, we use multiple video frames jointly to derive a global feature representation of a DHG video via average pooling on the temporal dimension, which can help reduce ambiguities among samples, resulting in more robust and discriminative identity features. For the second drawback, we use the PA (Persistence of Appearance) proposed in [16] to replace the optical flow as the behavior representation. The PA is computationally efficient and can be trained in an end-to-end fashion. For the third and fourth drawbacks, we adopt more advanced ResNet and AM-Softmax [15] as the backbone and loss function. Combining the above strategies, we redesign the two streams of the TS-CNN based on the work of Zhang *et al.* [16]. The enhanced architecture is shown in Fig. 1 where the spatial and temporal streams are mainly used to encode the physiological and behavioral information of hand gestures, respectively.

As shown in Fig. 1, we first obtain the input data for the two streams by a sparse sampling scheme proposed in the TSN [17]. Specifically, a DHG video is uniformly divided into N segments, each of which contains m (we set m as four in this paper) adjacent frames, resulting in N snippets $\{\{\mathbf{I}^1_{S_1}, \mathbf{I}^2_{S_1}, \mathbf{I}^3_{S_1}, \mathbf{I}^4_{S_1}\}, ..., \{\mathbf{I}^1_{S_N}, \mathbf{I}^2_{S_N}, \mathbf{I}^3_{S_N}, \mathbf{I}^4_{S_N}\}\}$. These snippets will be used to calculate behavior representation PA through the PA module in the temporal stream, and the first frame of each snippet $\{\mathbf{I}^1_{S_1}, ..., \mathbf{I}^1_{S_N}\}$ will be used as the inputs of the spatial stream.

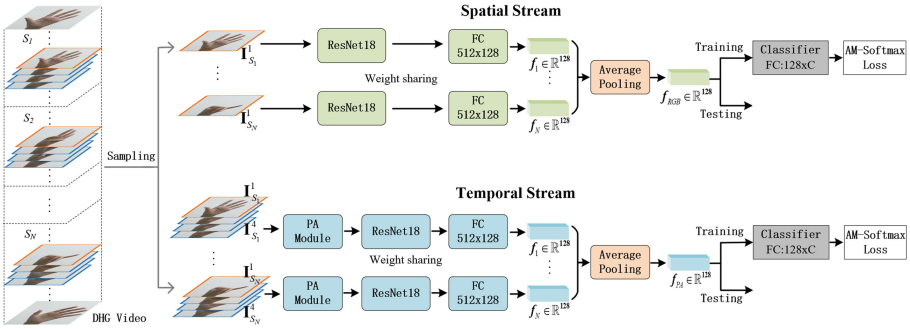


Fig. 1. Overview of our proposed improved two-stream CNNs.

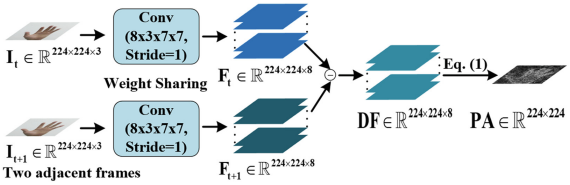


Fig. 2. Illustration of the PA module.

For the spatial stream, each frame is first encoded by a ResNet18 that outputs a 512-dim feature vector after the global average pooling (GAP) layer. This 512-dim feature vector is then transformed to a 128-dim feature vector by a fully-collected (FC) layer. The final physiological feature of a DHG in the spatial stream is represented as the average of all feature vectors. For the temporal stream, it works the same as the spatial stream except for the PA module. Figure 2 illustrates the working scheme of the PA module. It represents the motion between two adjacent frames as a single-channel PA map. Given two adjacent frames $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$, the PA module first obtains their corresponding low-level feature maps $\mathbf{F}_t$ and $\mathbf{F}_{t+1}$ through a lightweight convolution layer. Then, it calculates the difference between $\mathbf{F}_t$ and $\mathbf{F}_{t+1}$, obtaining the difference feature map **DF**. Finally, it transforms **DF** into a single-channel PA map as follows:

$$\boldsymbol{PA}(x,y) = \sqrt{\sum_{i=1}^{M}(\mathbf{DF}(x,y,i))^2} \tag{1}$$

where $\mathbf{DF}(x,y,i)$ represents the value of pixel (x,y) on the *i-th* channel of the **DF**, M is the channel numbers of **DF**, which is set to eight in this paper. For a snippet with m adjacent frames, the PA module calculates the PA map between every two adjacent frames and stacks the obtained $(m-1)$ PA maps together as the input to the ResNet18 in the temporal stream. Finally, both branches are trained under the supervision of the AM-Softmax loss [15]. In the testing stage, the cosine distance between the query feature and the registered DHG feature is calculated to measure their similarity. If the distance is smaller than the pre-defined threshold, the user is accepted, otherwise rejected.

2.3 Two-Stream Information Fusion

The raw RGB frames and the stacked PA maps contain physiological and behavioral information, respectively. The two kinds of information are complementary and can be merged to improve performance. Thus, it is significant to explore a reasonable fusion strategy for the two-stream information.

In this paper, we comprehensively investigate three mainstream fusion strategies, including early fusion, intermediate fusion, and late fusion, to address the fifth drawback to some extent. For the early fusion, the original RGB frames and stacked PA maps are concatenated together and then sent to the backbone for identity feature extraction. For the intermediate fusion, the information fusion is performed after the FC layer, and the final identity feature is the weighted sum of the two feature vectors, $\boldsymbol{f}_{RGB}$ and $\boldsymbol{f}_{PA}$, extracted from the spatial and temporal streams. For the late fusion, we first calculate the cosine distance of the two DHG features for each stream individually and then obtain the final distance by the weighted sum of the two distances.

3 Experiments

3.1 Dataset and Settings

We conduct extensive experiments on the large-scale SCUT-DHGA dataset [13], which is the only publicly available dynamic hand gesture authentication dataset that contains RGB videos, to demonstrate the effectiveness of our proposed method. The SCUT-DHGA dataset contains 29,160 DHG videos collected from 193 subjects. Videos from 143 subjects are divided into the training set, while videos from the other 50 subjects are divided into the test set. The videos in the test set are collected across two stages with an average interval of about one week, resulting in two test settings, *i.e.,* single-session and cross-session authentication. In the single-session authentication, the registered and query video both come from the first stage, while in the cross-session authentication, they come from two different stages.

In this paper, we focus on two main test protocols of the SCUT-DHGA dataset: *MG* and *UMG*. For the *MG* protocol, models are trained with videos from the six DHG types and are tested using videos from the same six DHG types. For the *UMG* protocol, the models are trained with videos from five DHG types and are tested using videos from the one remaining DHG type. Therefore, both the *MG* and *UMG* have six test settings, corresponding to six gesture types, respectively.

3.2 Implementation Details

We follow the common practice to initialize the backbone (ResNet18) using the weights pretrained on the ImageNet dataset. We also initialize the PA module using the weights pretrained on the Something-Something-V2 dataset. All networks are optimized by Adam. The weight decay factor and learning rate are set to 1e–7 and 1e–5, respectively. When training the spatial and temporal stream networks, we set the mini-batch size to 45 and 32, respectively. We adopt online data augmentations during training, including random rotation ($\pm 15^{\circ}$) and color jittering (± 0.3). We also dropout the DHG video

features, $\boldsymbol{f}_{RGB}$ and $\boldsymbol{f}_{PA}$, by a probability of 0.5. The spatial and temporal streams are trained for 50 and 200 epochs, respectively. Finally, all experiments are implemented using Pytorch on an Nvidia GTX1080Ti GPU.

Table 1. Comparisons between our proposed improved two-stream CNN and the baseline DHGA-net. The inputs to the spatial and temporal streams are the sampled RGB frames and PA maps, respectively.

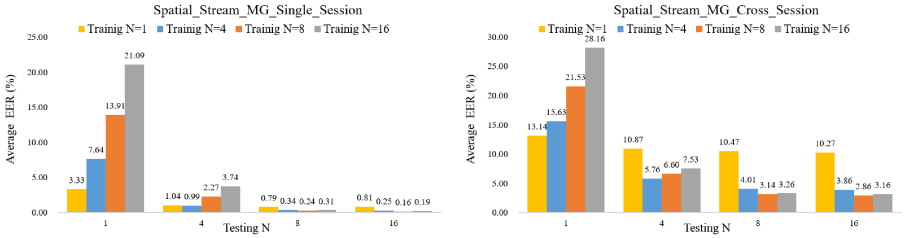
Setting	Single session EER(%)			Cross session EER (%)		
	DHGA-net	Spatial stream	Temporal stream	DHGA-net	Spatial stream	Temporal stream
UMG_g_1	2.53	**0.13**	**1.10**	13.40	**4.40**	**7.16**
UMG_g_2	2.84	**0.18**	**1.07**	13.87	**3.09**	**5.22**
UMG_g_3	2.00	**0.36**	**0.49**	10.84	**3.78**	**4.89**
UMG_g_4	1.82	**0.71**	**0.76**	8.20	**2.78**	**3.44**
UMG_g_5	2.04	**0.18**	**0.49**	10.96	**3.09**	**3.84**
UMG_g_6	2.36	**0.13**	**2.02**	13.60	**4.04**	**6.93**
Average	2.27	**0.28**	**0.99**	11.81	**3.53**	**5.25**

3.3 Performance of the Two Single Streams

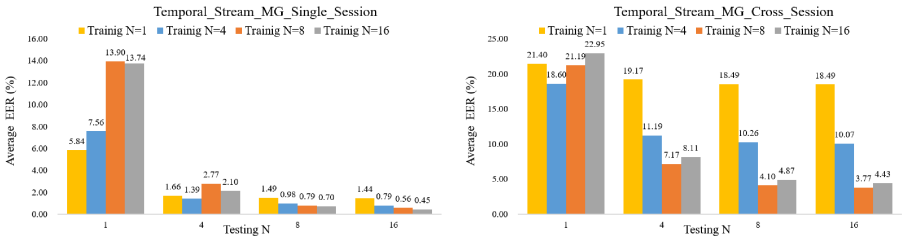
We first evaluate the performance of the spatial and temporal streams of our proposed improved two-stream CNN. The comparison with the baseline DHGA-net proposed for benchmarking the SCUT-DHGA dataset [13] is shown in Table 1. The results manifest that both spatial and temporal streams significantly outperform the DHGA-net. In the single session, the average EER of the DHGA-net is 2.27% under the *UMG*, while the spatial and temporal streams can achieve 0.28% and 0.99% average EERs, decreasing the average EERs by 1.99% and 1.28% respectively. In the cross session, the average EER of the DHGA-net is 11.81% under the *UMG*, while the spatial and temporal streams can achieve 3.53% and 5.25% average EERs, decreasing the average EERs by 8.28% and 6.56%, respectively. It can not only demonstrate the effectiveness of our improvement strategies but also can justify the feasibility of both physiological and behavioral characteristics for DHG authentication. Besides, the results also show that the performance of the spatial stream is better than that of the temporal stream. The reasons are two-fold. First, the behavioral characteristic understanding involves extracting fine-grained spatiotemporal features, which is more difficult than the physiological characteristic understanding. Second, the representation of behavioral characteristics (PA) is not optimal and needs further improvement.

3.4 Impact of Local Feature Number

To verify the impact of the local feature number (i.e., the number of segments, *N*, in Fig. 1) on the performance of the spatial and temporal streams, we conduct comprehensive comparison experiments under the *MG*. The comparison results are shown in Fig. 3. We can see that the results of both streams have the same trend. When *N* is fixed during training, the average EER will decrease to saturation with the increase of *N* in testing. This can prove that the combination of multiple local features is very crucial for authentication performance improvement, but too many local features will cause information redundancy and can not contribute to a significant boost. When the *N* in testing



(a) Comparison results on the spatial stream



(b) Comparison results on the temporal stream

Fig. 3. Comparison results of different local feature number.

is small (for example, less than or equal to 4) and less than the N in training, the average EER becomes significantly larger with the increase of N during training. The reason is that the training process makes the networks only discriminative for features that combine enough local information. When the N in testing is relatively large (for example, 8 or 16) and greater than the N in training, the N during training will have a significant impact on the authentication performance. For example, setting N to 1 during training will cause a considerable EER increase. It is because minimal local features have very limited discriminative information, making networks hard to train. These results clearly demonstrate the effectiveness of the first improvement strategy for the first drawback of the TS-CNN.

3.5 Impact of Behavior Representation

The results in Table 1 demonstrate the effectiveness of the PA (the temporal stream). Here, we compare the PA with the optical flow to further demonstrate its superiority in behavior representation by replacing the optical flow with PA maps in the temporal stream. In this paper, we adopt the same method as the TS-CNN for optical flow extraction offline. The comparison results under the *MG* setting are shown in Table 2. The results show that the PA can achieve a lower EER than the optical flow for most DHG categories, and the PA's average EER is lower than the optical flow by 0.50% and 0.76% in single session and cross session. We attribute the success of the PA to its ability to adaptively generate decent behavior representations for hand gesture authentication. In addition, it is important to mention that the PA can be derived online and is therefore more suitable for deployment in authentication systems than the optical flow.

Table 2. Comparisons between the optical flow and PA.

Setting	Single session EER (%)		Cross session EER (%)	
	Optical Flow	PA	Optical Flow	PA
MG_g_1	1.56	**0.71**	7.36	**5.33**
MG_g_2	1.82	**1.40**	5.71	**4.42**
MG_g_3	0.71	**0.53**	**3.87**	3.98
MG_g_4	1.36	**0.71**	4.62	**3.73**
MG_g_5	0.84	**0.56**	3.22	**3.07**
MG_g_6	1.44	**0.80**	4.36	**4.04**
Average	1.29	**0.79**	4.86	**4.10**

Table 3. Comparisons between two different backbones: ResNet18 and AlexNet.

Setting		EER (%)			
		Spatial stream		Temporal stream	
		AlexNet	**ResNet18**	AlexNet	**ResNet18**
Single session	MG_g_1	1.42	**0.18**	3.33	**0.71**
	MG_g_2	0.89	**0.31**	2.80	**1.40**
	MG_g_3	0.93	**0.09**	2.44	**0.53**
	MG_g_4	1.29	**0.53**	3.24	**0.71**
	MG_g_5	0.76	**0.18**	1.91	**0.56**
	MG_g_6	0.84	**0.18**	3.84	**0.80**
	Average	1.02	**0.24**	2.93	**0.79**
Cross session	MG_g_1	7.78	**4.09**	12.64	**5.33**
	MG_g_2	6.64	**3.27**	11.38	**4.42**
	MG_g_3	7.04	**2.93**	10.82	**3.98**
	MG_g_4	5.93	**2.58**	12.42	**3.73**
	MG_g_5	6.73	**2.49**	9.56	**3.07**
	MG_g_6	7.09	**3.47**	11.73	**4.04**
	Average	6.87	**3.14**	11.43	**4.10**

Table 4. Comparisons among three different loss functions: Softmax, Softmax+Center, AM-Softmax.

Setting			EER (%)		
			Softmax	Softmax+Center	**AM-Softmax**
Spatial stream	Single session	MG_g_1	0.91	0.58	**0.18**
		MG_g_2	0.84	0.62	**0.31**
		MG_g_3	0.84	0.36	**0.09**
		MG_g_4	1.16	**0.27**	0.53
		MG_g_5	0.89	0.36	**0.18**
		MG_g_6	0.87	0.31	**0.18**
		Average	0.92	0.41	**0.24**
	Cross session	MG_g_1	10.13	8.49	**4.09**
		MG_g_2	14.13	10.38	**3.27**
		MG_g_3	10.24	5.51	**2.93**
		MG_g_4	9.76	5.53	**2.58**
		MG_g_5	9.44	4.71	**2.49**
		MG_g_6	9.98	6.24	**3.47**
		Average	10.61	6.81	**3.14**
Temporal stream	Single session	MG_g_1	0.73	**0.62**	0.71
		MG_g_2	1.64	**1.07**	1.40
		MG_g_3	1.04	**0.44**	0.53
		MG_g_4	0.78	**0.58**	0.71
		MG_g_5	0.73	**0.36**	0.56
		MG_g_6	1.47	1.20	**0.80**
		Average	1.07	**0.71**	0.79
	Cross session	MG_g_1	12.27	10.20	**5.33**
		MG_g_2	10.16	9.18	**4.42**
		MG_g_3	9.31	6.60	**3.98**
		MG_g_4	7.56	4.98	**3.73**
		MG_g_5	5.07	4.33	**3.07**
		MG_g_6	8.96	8.36	**4.04**
		Average	8.89	7.27	**4.10**

3.6 Impact of Backbone

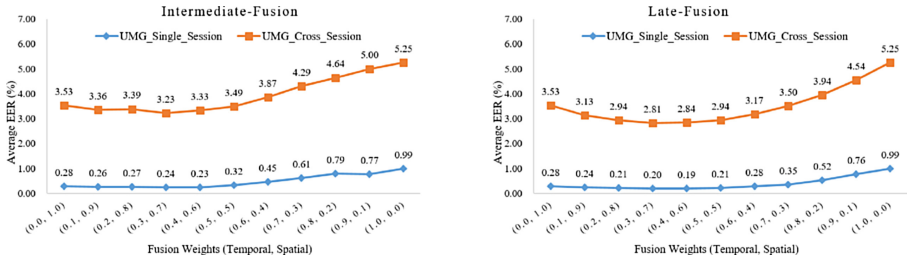
The backbone has a significant impact on image feature extraction. In this section, we compare our adopted RestNet18 with the AlexNet used in the TS-CNN. The comparison results under the *MG* are listed in Table 3. It indicates that the ResNet18 can consistently outperform the AlexNet by large margins for both spatial and temporal streams due to its more rational design and more powerful non-linear representation capability in hand gesture authentication.

3.7 Impact of Loss Function

The loss function is vital for model optimization. In this section, we compare three loss functions, including Softmax, Softmax plus Center [18], and AM-Softmax. The comparison results under the *MG* are shown in Table 4. It shows that both the Softmax plus Center and AM-Softmax can significantly outperform the Softmax, which proves the importance of adding constraints to decrease intra-class variation. Moreover, our adopted AM-Softmax loss function also significantly outperforms Softmax plus Center in the cross-session. Overall, our adopted AM-Softmax performs best.

3.8 Comparisons Among Different Two-Stream Information Fusion Methods

We compare three different two-stream information fusion methods, including early fusion, intermediate fusion, and late fusion, under the *UMG*. We first verify the per-



(a) (b)

Fig. 4. Results of intermediate-fusion and late-fusion with different fusion weights.

formance of the intermediate-fusion and late-fusion with different fusion weights. The results in Fig. 4 show that setting the weight of the temporal stream to be smaller than the weight of the spatial stream is good for both intermediate fusion and late fusion. This is in line with our expectations because the performance of the spatial stream is better than that of the temporal stream, as shown in Table 1. Thus, it is very reasonable to give the spatial stream a higher weight. Finally, when the weights are set to 0.3 and 0.7 for the temporal and spatial stream, both intermediate-fusion and late-fusion can achieve relatively good results. Thus, we adopt this weight setting for the intermediate-fusion and late-fusion in subsequent experiments to compare with the early-fusion, and the results are listed in Table 5. We can find that the late-fusion can perform best because it can make a comprehensive evaluation according to the similarities of physiological and behavioral characteristics by setting an appropriate fusion weight based on their actual performance.

Table 5. Comparisons among three different two-stream information fusion methods.

Setting	Single session EER (%)			Cross session EER (%)		
	Early-fusion	Intermediate-fusion	Late-fusion	Early-fusion	Intermediate-fusion	Late-fusion
UMG_g_1	0.51	0.18	**0.13**	3.71	4.02	**3.47**
UMG_g_2	0.69	0.18	**0.13**	4.29	2.84	**2.44**
UMG_g_3	0.89	0.22	**0.09**	**2.67**	3.98	3.51
UMG_g_4	0.73	**0.53**	0.58	2.76	2.27	**1.82**
UMG_g_5	0.31	0.13	**0.13**	2.53	2.33	**1.98**
UMG_g_6	0.53	0.22	**0.13**	4.20	3.91	**3.67**
Average	0.61	0.24	**0.20**	3.36	3.23	**2.81**

4 Conclusion

In this paper, we first analyze the authentication-oriented TS-CNN in depth, expecting to fully unleash the potential of two-stream CNNs in DHG authentication. We then propose the improved two-stream CNNs for DHG authentication by enhancing the TS-CNN from five aspects. The extensive results on the SCUT-DHGA dataset demonstrate the effectiveness of the introduced methods.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 61976095.

References

1. Aumi, M.T.I., Kratz, S.G.: Airauth: evaluating in-air hand gestures for authentication. In: MobileHCI 2014 (2014)
2. Sun, Z., Wang, Y., Qu, G., Zhou, Z.: A 3-d hand gesture signature based biometric authentication system for smartphones. Secur. Commun. Netw. **9**, 1359–1373 (2016)
3. Karita, S., Nakamura, K., Kono, K., Ito, Y., Babaguchi, N.: Owner authentication for mobile devices using motion gestures based on multi-owner template update. In: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6 (2015)
4. Sae-Bae, N., Memon, N., Isbister, K., Ahmed, K.: Multitouch gesture-based authentication. IEEE Trans. Inf. Forensics Secur. **9**, 568–582 (2014)
5. Shen, C., Zhang, Y., Guan, X., Maxion, R.: Performance analysis of touch-interaction behavior for active smartphone authentication. IEEE Trans. Inf. Forensics Secur. **11**, 498–513 (2016)
6. Tian, J., Qu, C., Xu, W., Wang, S.: Kinwrite: handwriting-based authentication using kinect. In: NDSS (2013)
7. Wu, J., Christianson, J., Konrad, J., Ishwar, P.: Leveraging shape and depth in user authentication from in-air hand gestures. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 3195–3199 (2015)
8. Nugrahaningsih, N., Porta, M., Scarpello, G.: A hand gesture approach to biometrics. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 51–58. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_7
9. Wang, X., Tanaka, J.: Gesid: 3D gesture authentication based on depth camera and one-class classification. Sensors (Basel, Switzerland) **18**, 3265 (2018)
10. Gu, J., Wang, Z., Ouyang, W., Zhang, W., Li, J., Zhuo, L.: 3D hand pose estimation with disentangled cross-modal latent space. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 380–389 (2020)
11. Wu, J., Ishwar, P., Konrad, J.: Two-stream cnns for gesture-based verification and identification: Learning user style. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 110–118 (2016)
12. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
13. Liu, C., Yang, Y., Liu, X., Fang, L., Kang, W.: Dynamic-hand-gesture authentication dataset and benchmark. IEEE Trans. Inf. Forensics Secur. **16**, 1550–1562 (2021)
14. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733 (2017)
15. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**, 926–930 (2018)
16. Zhang, C., Zou, Y., Chen, G., Gan, L.: Pan: persistent appearance network with an efficient motion cue for fast action recognition. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 500–509 (2019)
17. Wang, L., Xiong, Y., Zhe Wang, Yu., Qiao, D.L., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2740–2755 (2019)
18. Wen, Y., Zhang, K., Li, Z., Qiao, Yu.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31