J Inf Process Syst, Vol.16, No.5, pp.1145~1157, October 2020
https://doi.org/10.3745/JIPS.04.0190

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

# Classroom Roll-Call System Based on ResNet Networks

Jinlong Zhu*, Fanhua Yu*, Guangjie Liu*, Mingyu Sun*, Dong Zhao*,
Qingtian Geng*, and Jinbo Su*

**Abstract**

A convolution neural networks (CNNs) has demonstrated outstanding performance compared to other algorithms in the field of face recognition. Regarding the over-fitting problem of CNN, researchers have proposed a residual network to ease the training for recognition accuracy improvement. In this study, a novel face recognition model based on game theory for call-over in the classroom was proposed. In the proposed scheme, an image with multiple faces was used as input, and the residual network identified each face with a confidence score to form a list of student identities. Face tracking of the same identity or low confidence were determined to be the optimisation objective, with the game participants set formed from the student identity list. Game theory optimises the authentication strategy according to the confidence value and identity set to improve recognition accuracy. We observed that there exists an optimal mapping relation between face and identity to avoid multiple faces associated with one identity in the proposed scheme and that the proposed game-based scheme can reduce the error rate, as compared to the existing schemes with deeper neural network.

**Keywords**

Face Recognition, Game, ResNet Networks

## 1. Introduction

The main focus of face recognition is to improve accuracy [1]. To date, the face recognition algorithm based on convolutional neural network (CNN) [2], keeps refreshing the recognition accuracy, which includes LeNet [3], AlexNet [4], VGGNet [5], ZFNet [6], GoogLeNet [7], ResNet [8], and DenseNet [9].

The LeNet model is one of the best networks for the elementary convolution neural network, including the convolution, pooling, and fully connected layers. Since then, the basic architecture of CNN has been defined as a convolution layer, pooling layer, and fully connected layer [10,11]. The AlexNet model applies the basic principles of CNN to a deep and wide network. It successfully applies the trick in CNN to improve performance, such as ReLU, dropout, and LRN (local response normalization). Meanwhile, AlexNet also relies on a GPU for operation acceleration. Researchers change the network structure or convolution kernel to improve the algorithm performance, such as ZFNet, VGGNet, and GoogleLeNet.

The ResNet proposed by He et al. [8] in 2015 outperformed all other algorithms on ISLVRC and COCO datasets and won the ILSVRC15 championship. The ResNet solves the network degradation problem, which refers to the issue of saturating or declining network accuracy owing to increasing network depth.



Manuscript received June 26, 2019; first revision August 19, 2019; December 21, 2019.

**Corresponding Author:** Fanhua Yu (258362073@qq.com)

* Dept. of Computer Science and Technology, Changchun Normal University, Changchun, China (zhujinlong19840913@126.com, 258362073@qq.com, 756058599@qq.com, sunmingyu370@163.com, zhaodong@126.com, 330561812@qq.com)

As inspired by ResNet, DenseNet was proposed in 2017, which was a feature map directly merged from different layers. However, it consumes a significant amount of memory in training due to its poor implementation. AlexNet, ZFNet, and VGGNet have inferior performance than GoogleLeNet and ResNet. Meanwhile, ResNet designed a residual module to solve the problem of vanishing gradient to allow training deeper network.

In this study, we propose a face recognition model based on ResNet and game for roll-call in the classroom. The model recognises all face identities in an image, with the network output being a list of identities with confidence scores to recognise a face image. We take identities with confidence values greater than 0.6 as the attendance set, and the others as the undetermined set. Game theory obtains the Nash equilibrium to construct the optimal identity combination and to determine the attendance status of students in the uncertain identity set.

The rest of this paper is organised as follows. A brief ResNet network overview is presented in Section 2. Section 3 describes the face recognition model based on the ResNet network and game for roll-call in the classroom. In Section 4, the experimental results illustrate how the model benefits in terms of face recognition accuracy rate. Finally, Section 5 concludes the paper.

## 2. ResNet Network

In mathematical statistics, residual refers to the difference between the actual observed value and the estimated value (fitting value). Researchers have introduced the concept of residual into convolutional neural network to construct the residual network. The main characteristic is its easy network optimisation [12] and ability to improve the accuracy by increasing the depth. The internal residual block uses a jump connection, which alleviates the vanishing gradient problem [8,13] caused by the increasing network depth [14].

If the later layers are identity mapping, the model will degenerate into a shallow network. To address this problem, the residual network learns the identity mapping function to allow some layers *to fit a potential identity mapping function* $H(x) = x$. The residual network design structure is $H(x) = F(x) + x$, as shown in Fig. 1. We can convert it into learning a residual function, $F(x) = H(x) - x$. If the function satisfies $F(x) = 0$, it forms an identity mapping $H(x) = x$. The mapping of the network with residual is more sensitive to the output change, and greater output change adjustment corresponds to greater weight, leading to better training effect. The idea of residuals is to remove the same main body, thereby highlighting small changes.

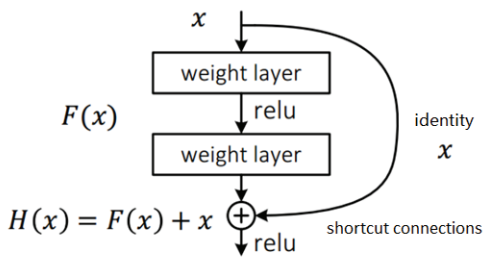



**Fig. 1.** Structure of the residual network.

## 3. ResNet-Based Roll-Call System with Game Theory

The proposed method uses the Inception_Reset2 network for training face sample data (Labeled Faces in the Wild [LFW]) to obtain a recognition model. The Inception_Reset2 network structure is shown in Fig. 1. The structures of Inception_Resnet_A, Inception_Resnet_B, and Inception_Resnet_C in the Inception_Reset2 structure are illustrated in Figs. 2, 3, and 4, respectively. The classical inception module

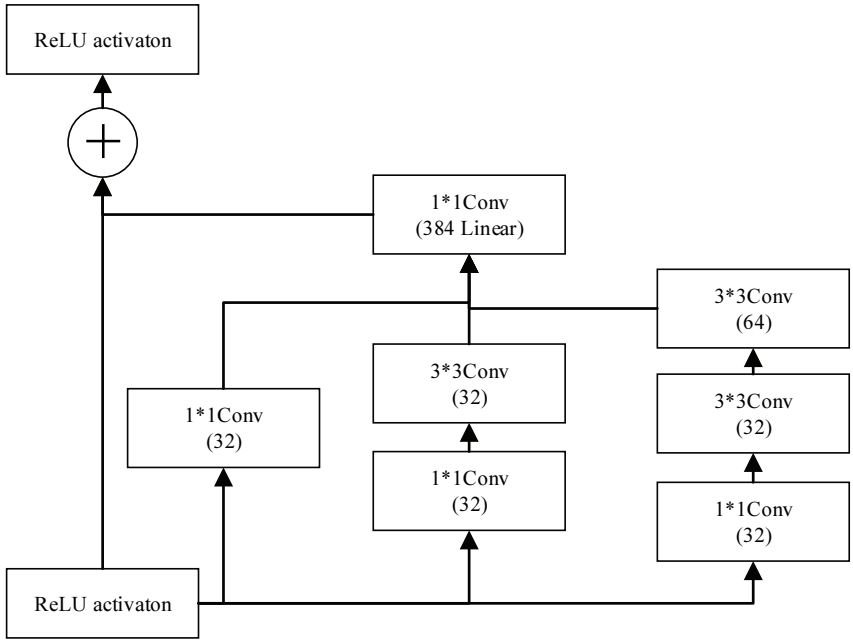



**Fig. 2.** Structure of Inception_Resnet_A.

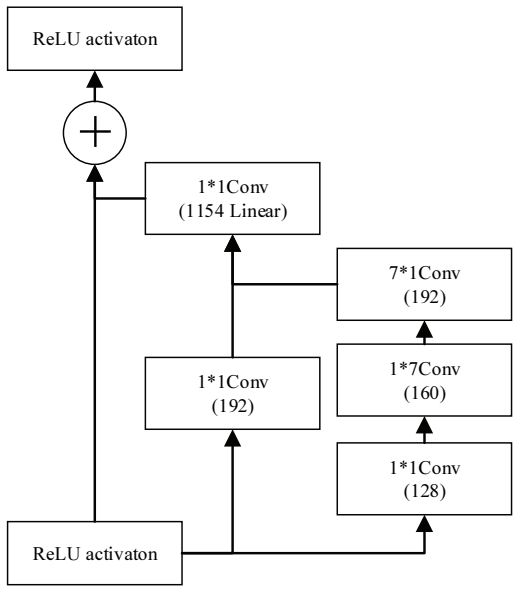



**Fig. 3.** Structure of Inception_Resnet_B.

has three filters that perform convolution operations on the input, which are different in sizes: 1×1, 3×3, and 5×5. The output of all sublayers is cascaded and transmitted to the next inception module. The computational cost is more expensive on A 5×5 convolution and more than 2.78 times [15] less on a 3×3 convolution. Thus, superimposing two 3×3 convolutions can improve the network performance. An additional 1×1 convolution layer is added to the 3×3 convolution layer to limit the number of input channels in Fig. 2. A 3×3 convolution is equivalent to a 1×3 convolution and a 3×1 convolution. The 1×3 convolution and 3×1 convolution also found that the cost of this method was 33% lower than that of a single 3×3 convolution in Figs. 3 and 4. Thus, the network structure can reduce the parameters and the computation of the model without affecting the results.

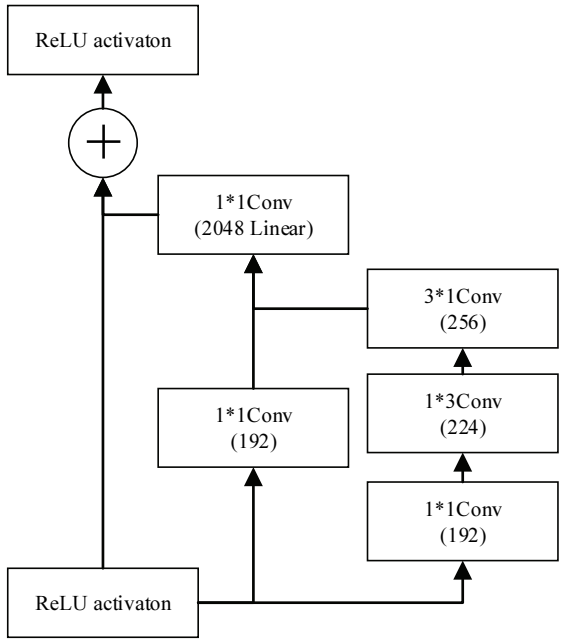



**Fig. 4.** Structure of Inception_Resnet_C.

As shown in Fig. 5, the process includes the following steps. We integrate the face images of students into LFW as sample data. The ResNet network (Inception_Resnet_v2) trains the sample data to generate a face recognition model, which serves as our feature extractor. The residual network combines with support vector machine (SVM) to recognise face identity with corresponding confidence value. SVM encodes the feature vector to generate different face classifiers according to different classes. The face classifier generates a mapping list of identity and confidence to identify a face in the attendance picture. The format of a mapping list record is **<**$iden_i$,$conf_i$**>**, where $iden_i$ represents the $i$-th identity in the class, while $conf_i$ represents the corresponding confidence value. We take the maximum confidence value $C_m$ as the identity in the confidence list. If $C_m$>0.6, the identity is added to the attendance set $I_a$; otherwise, it will add the mapping list to the uncertain identity set $I_u$. If the maximum confidence values of different face images are identical, then their mapping list is added to $I_u$. The data format in $I_u$ is <$id_f$, $iden_i$, $conf_i$>, where $id_f$ indicates the face image number of the record in the mapping list.

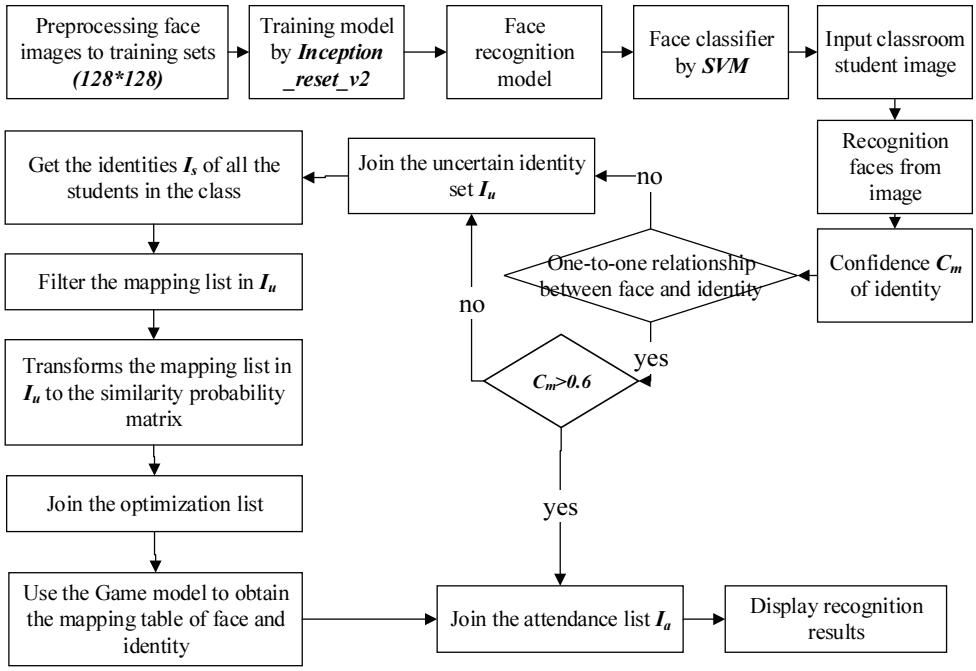



**Fig. 5.** Flow diagram of the algorithm.

Suppose that all the student identities in the class constitute the set $I_s$. Then, $I_s=I_a+I_c$, where $I_c$ is the identity set of absent students. We filter the mapping list in $I_u$ and satisfy that the identity of the record in the mapping list belongs to $I_c$. The proposed method transforms the mapping list in $I_u$ into a similarity probability matrix, whose horizontal axis indicates the student identity in $I_c$ and the vertical axis corresponds to the face image in $I_u$. The values stored in this matrix are the confidence values of the map of faces and identities.

Nash equilibrium is a strategic combination of all participants, in which each person's strategy is the best response to the strategy of others. The Nash equilibrium is a type of equilibrium achieved in the continuous game.

The game theory model includes participants, strategies, and payoffs [16]. As a general case, we consider the similarity probability matrix as the payoff matrix to obtain Nash equilibrium. The payoff matrix can describe the probability of face recognition when choosing a strategy. This can help us determine the optimal strategy combination as the face recognition result to improve accuracy. The payoff function is expressed as follows:

$$U = \sum_{i \in n} p(x_{ij}) \quad , i \neq j \, , j \in m. \tag{1}$$

In Formula (1), $n$ denotes the number of faces and $m$ denotes the number of identities in the similarity probability matrix. Further, $p(x_{ij})$ is the confidence value with a map of face $i$ and identity $j$. The maximum value of the payoff function $U$ is the Nash equilibrium point to generate the mapping table of the face and identity. The mapping relation in the mapping table is that the recognition result of a face image is added to the attendance list.

# 4. Performance Analysis

The core idea of Inception_v1 [7] uses 1×1, 3×3, and 5×5 convolution layers, instead of some large convolution layers of GoogleNet, which can significantly reduce the number of weight parameters. Inception_v2 [17] adds batch normalisation to input, which facilitates training convergence and learning more efficiently. We choose Inception_v2 as the contrast algorithm because its performance is better than that of Inception_v1. Inception_v3 transforms convolutions 7×7 in GoogleNet into two layers of 1×7 and 7×1 in series, as well as 3×3 into 1×3 and 3×1, which speeds up the calculation, increases the nonlinearity of the network, and reduces the probability of over-fitting. Inception_v4 mainly uses the residual connection to improve the structure of Inception_v3. The representative network is Inception_Resnet_v1, Inception_Resnet_v2, and Inception_v4. Inception_v4 adds the ResNet structure to Inception, and a node can skip from some nodes directly connected to the following nodes, while the residuals follow the past one. Combining the Inception module with residual connection, Inception_Resnet_v1 and Inception_Resnet_v2 are proposed, which lead to faster training convergence and higher accuracy. A deeper version of Inception_v4 is designed; its effect is comparable to that of Inception_Resnet_v2.

Finally, we present some comparisons between various versions of Inception and Inception_Resnet. The models Inception_v2 and Inception_v3 are deep convolutional networks, while Inception_Resnet_v1 and Inception_Resnet_v2 are Inception-style networks that utilise residual connections instead of filter concatenation. Inception_v4 is a pure inception variant without residual connections with roughly the same recognition performance as Inception_Resnet_v2 [18].

The residual network training model takes LFW as the sample database, whereas the SVM method takes the face photos of all the students in the class as training samples. A comparison of the error rates of the algorithms in top 1 and top 5 is presented in Table 1. It can be observed that the Inception_Reset_v2 demonstrates better recognition performance than the other algorithms. Therefore, we choose Inception_Reset_v2 as our residual network structure for face identification.

**Table 1.** Error rate comparison of the network model

| Network | Error (%) | |
|---|---|---|
| | Top_1 | Top_5 |
| Inception_v2 | 18.8 | 4.5 |
| Inception_v3 | 18.7 | 4.4 |
| Inception_Reset_v1 | 17.6 | 3.8 |
| Inception_Reset_v2 | 17.5 | 3.7 |

Face detection aims to locate the face in an image, mark its coordinates, and crop it out to be the input image. As shown in Fig. 6, there are three results for the residual network in face recognition:

1. Recognition identity is correct ($C_m$>0.6).
2. Recognition identity is a mistake, which will lead to a mapping relationship between an identity and multiple faces or a wrong identity relationship in face mapping. The main reasons for this are as follows:
   - Large distance between the face and camera leads to a blurred image.
   - Facial region is occluded by other objects in the test image.
   - Hair occludes most facial regions.

3. The algorithm cannot recognise the face area in the test image because the face is far away from the camera.

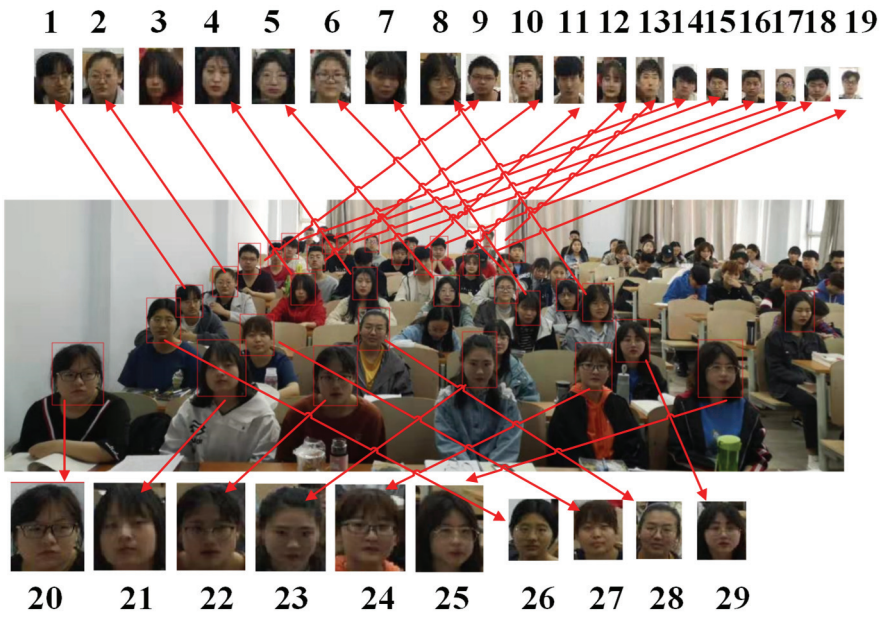



**Fig. 6.** Test image.

This algorithm detects the identity of all face images as an attendance list, which contains correct identities *ci* and error identities *ei*. The recognition error rate *ER* is a quotient of *the wrong identity number ei* and attendance list *al* in Formula (2).

$$ER = ei/al \tag{2}$$

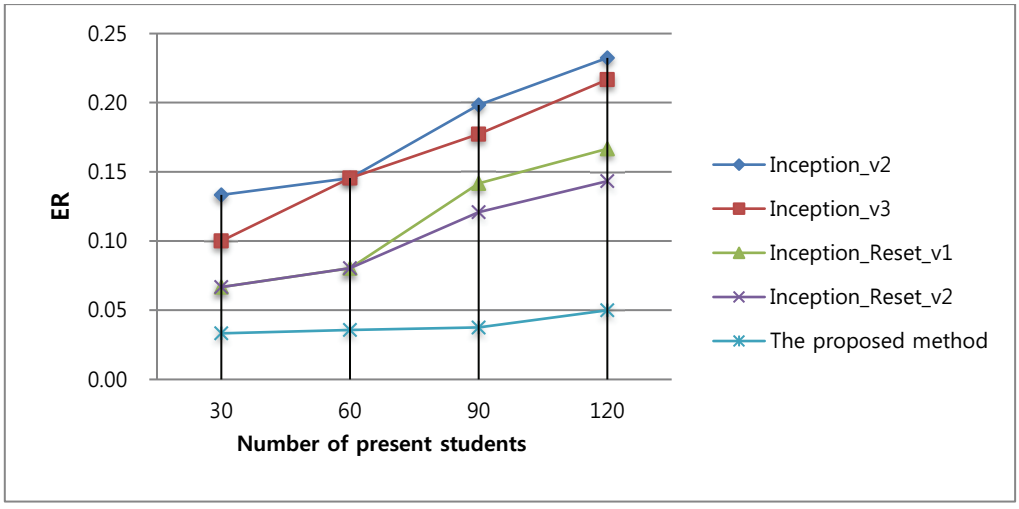



**Fig. 7.** Error rate diagram of different numbers of present students.

Fig. 7 shows that the error rate of our proposed algorithm is lower than that of the other methods. The main reason is that other algorithms map multiple faces to the same identity. Our method is more advantageous in recognising all student faces in an image, compared with other methods only for single face. It optimises the similarity probability matrix to generate the optimal mapping relationship between the faces and identities. The optimal mapping relationship is the product of maximising the confidence of all identities. It can not only ensure the one-to-one relationship between face and identity but also improves the accuracy of the mapping relationship.

The unrecognised rate *UR* is a quotient between the number (*s-ci-ei*) of unrecognised faces and the total number of student faces in Formula (3).

$$UR = (s - ci - ei)/s \tag{3}$$

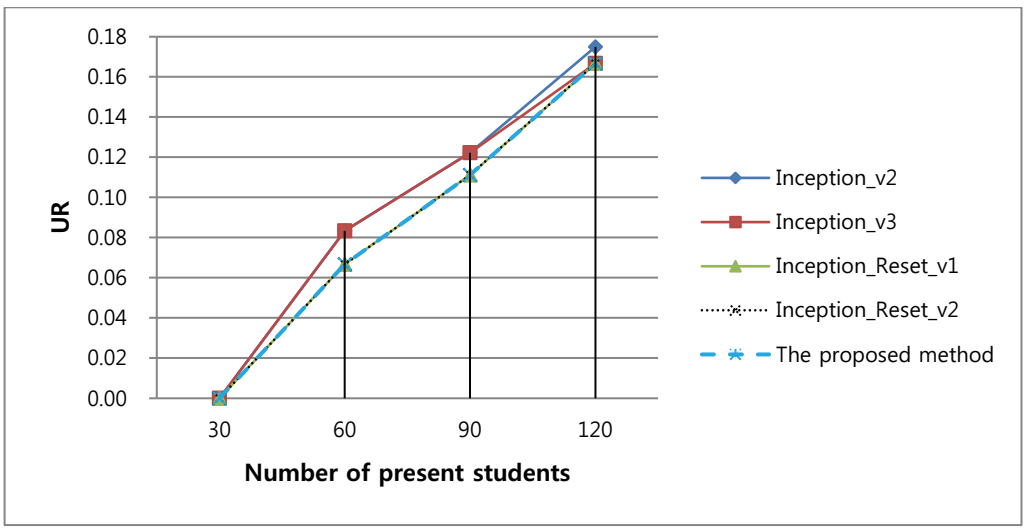



**Fig. 8.** Unrecognized rate diagram of different numbers of student attendance.

The number of unrecognised faces in the test image is given in Fig. 8. The confidence value of the mapping relationship between the face and identity generated by the Resnet network is the optimisation objective of our method. Therefore, the recognition performance of our proposed method is the same as Inception_Reset_v1 and Inception_Reset_v2 in terms of the number of unrecognised faces.

The ER and UR effects of absenteeism on 60 students in the classroom are demonstrated in Figs. 9 and 10. Meanwhile, a comparison of the error rate and the unrecognised rate for 120 students are illustrated in Figs. 11 and 12. We compared the proposed scheme with other schemes in terms of error and unrecognised rates. The analytical results revealed that our method solves the problem of multiple faces mapping one identity and optimises the global relationship between face and identity, which improves the algorithm accuracy. It can be observed from Table 2 that our proposed method has the highest average accuracy compared with the other methods.

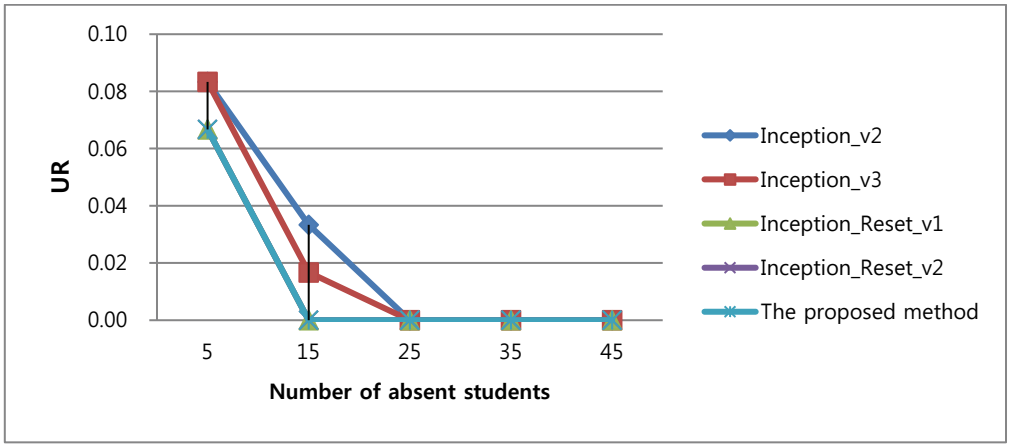



**Fig. 9.** Contrast chart of unrecognized rate under 60 students with different numbers of absent students.

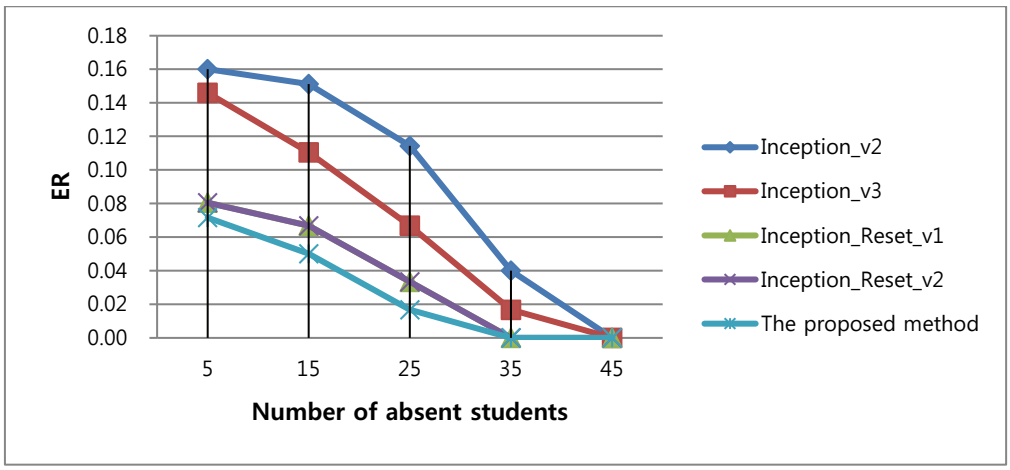



**Fig. 10.** Contrast chart of error rate under 60 students with different numbers of absent students.

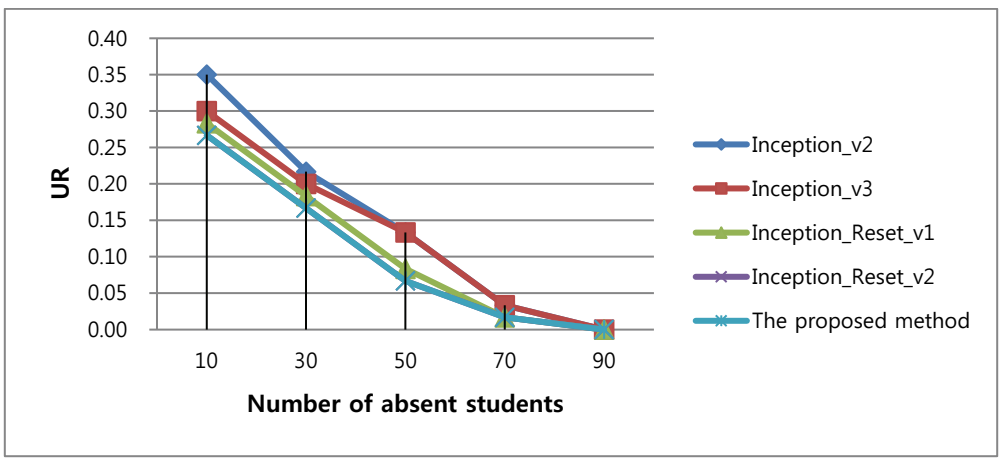



**Fig. 11.** Unrecognised rate under 120 students with different numbers of absent students.

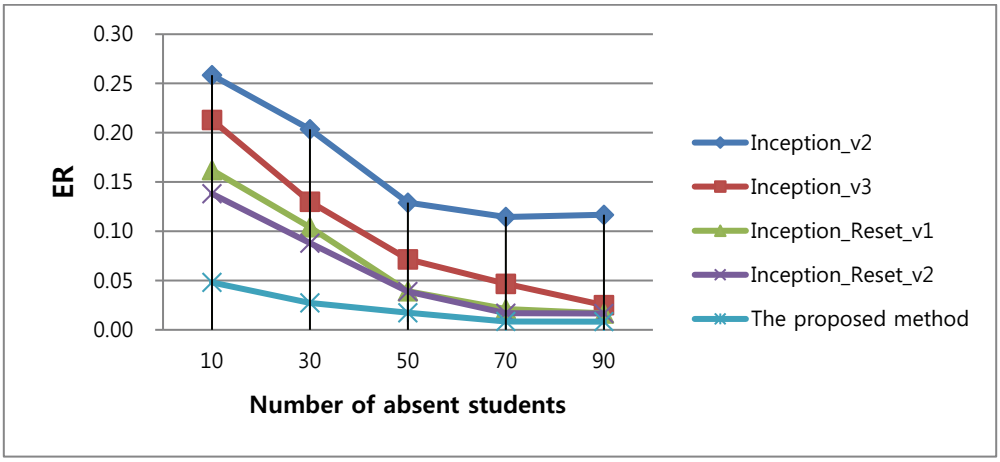



**Fig. 12.** Contrast chart of error rate under 120 students with different numbers of absent students.

**Table 2.** Error rate comparison table of the classroom roll-call system

| No | Inception_v2 | Inception_v3 | Inception_Reset_v1 | Inception_Reset_v2 | Proposed method |
|---|---|---|---|---|---|
| 1 | 0.35 | 0.3 | 0.283333 | 0.266667 | 0.266667 |
| 2 | 0.216667 | 0.2 | 0.183333 | 0.166667 | 0.166667 |
| 3 | 0.133333 | 0.133333 | 0.083333 | 0.066667 | 0.066667 |
| 4 | 0.033333 | 0.033333 | 0.016667 | 0.016667 | 0.016667 |
| 5 | 0.258427 | 0.212958 | 0.162148 | 0.138036 | 0.048141 |
| 6 | 0.203463 | 0.129874 | 0.1041 | 0.087962 | 0.027295 |
| 7 | 0.129032 | 0.071511 | 0.039155 | 0.038806 | 0.017247 |
| 8 | 0.114583 | 0.046655 | 0.021017 | 0.01681 | 0.008405 |
| 9 | 0.116667 | 0.025024 | 0.01667 | 0.016669 | 0.008334 |
| 10 | 0.16 | 0.145879 | 0.080567 | 0.080473 | 0.071531 |
| 11 | 0.151163 | 0.110452 | 0.06679 | 0.066741 | 0.050056 |
| 12 | 0.114286 | 0.066794 | 0.03337 | 0.033352 | 0.016676 |
| Avg | 0.16508 | 0.122985 | 0.090874 | 0.08296 | 0.063696 |

## 5. Conclusion

In this study, we proposed a face recognition model based on ResNet and game theory for roll-call in the classroom. The Inception_Reset_v2 is combined with game theory to optimise the mapping relationship between face and identity to improve recognition accuracy. The main advantage of this scheme is that the accuracy is significantly improved compared with other networks in the recognition of multiple faces in a picture. In addition, game theory optimises the similarity probability matrix to find the optimal mapping relationship of faces and identities to solve multiple faces mapping the same identity. Numerical analysis revealed that the proposed scheme can provide better recognition accuracy than the existing scheme. This fact proves that the combination of neural networks and multi-objective optimisation algorithms can improve the application advantages in the production environment. The roll-call system requires 100% accuracy of face recognition. Our future work will aim to improve the

identification accuracy of the residual network. The improved method includes two steps: one is to improve the structure of the residual network, and the other is to investigate other CNNs.

## Acknowledgement

This paper is funded by National Natural Science Foundation of China (No. 61604019), science and technology development project of Jinlin Province (No. 20180201086SF), Education Department of Jilin Province (No. JJKH20181181KJ and JJKH20181165KJ), Jilin Provincial Development and Reform Commission (No. 2017C031-2), and Research Projects of Jilin Development and Reform Commission (No. 2019C039-1).

## References

[1] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *Image and Vision Computing*, vol. 79, pp. 35-48, 2018.

[2] X. X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," *Pattern Recognition*, vol. 45, no. 4, pp. 1318-1325, 2012.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, 2015.

[6] D. Matthew and R. Fergus, "Visualizing and understanding convolutional neural networks," in *Computer Vision – ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 818-833.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 1-9.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Las Vegas, NV, 2016, pp. 770-778.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 1243-1252.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 815-823.

[11] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1891-1898.

[12] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 1931-1939.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: integrated recognition, localization and detection using convolutional networks," in *Proceedings of the 2nd International Conference on Learning Representations*, Banff, Canada, 2014.

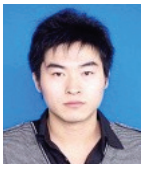
[14] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: a dataset for recognising faces across pose and age," in *Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, Xi'an, China, 2018, pp. 67-74.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 2818-2826.

[16] M. Mavronicolas and V. G. Papadopoulou, Algorithmic Game Theory. Heidelberg, Germany: Springer, 2009. https://doi.org/10.1007/978-3-642-04645-2

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 2818-2826.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," 2016 [Online]. Available: https://arxiv.org/abs/1602.07261.



**Jinlong Zhu** **https://orcid.org/0000-0002-2341-0542**

He was born in 1984. He received the B.E. and M.S. degrees in Computer Science and Technology from Jilin University, China, in 2003 and 2007, respectively. He is a professor at College of Computer Science and Technology, Changchun Normal University, Jilin, China. His research interest include digital image processing, virtual reality technique, computer algorithms and simulation, evacuation planning, image processing, and 3D modelling.

**Fanhua Yu** **https://orcid.org/0000-0002-0775-111X**

He received D.S. degree in School of Computer Science and Engineering from Jilin University in 2003. He is a professor at College of Computer Science and Technology, Changchun Normal University, Jilin, China. His current research interests include artificial neural network and fuzzy system.

**Guangjie Liu** **https://orcid.org/0000-0002-9232-9211**

He received D.S. degree in School of Computer Science and Technology from Jilin University in 2003. He is a professor at College of Computer Science and Technology, Changchun Normal University, Jilin, China, and the president of the School of Computer Science and Technology. His major research interests include computer graphics and computer image processing.

**Mingyu Sun** **https://orcid.org/0000-0002-7053-5524**

He received D.S. degree in School of Computer Science and Technology from Jilin University in 2016. His major research interests include computer graphics and computer image processing.

**Dong Zhao** **https://orcid.org/0000-0002-8313-8835**

He is an associate professor, School of computer science and technology, Changchun Normal University, Jilin, China. He received M.S. degree in computer science and technology, Changchun University of Science and Technology in 2008. His research interests include intelligent information system and embedded technology.

**Qingtian Geng** **https://orcid.org/0000-0003-3764-3037**

He is currently an associate professor in College of Computer Science and Technology, Changchun Normal University, Changchun, China. He received his B.S. and M.S. degrees in Computer Science and Technology from Jilin University of Technology and Jilin University in 1996 and 2005, respectively. He received his Ph.D. degree in Computer Science and Technology from Jilin University in 2016. His current research fields include image processing and pattern recognition.

**Jinbo Su** **https://orcid.org/0000-0001-8828-1967**

He is currently an undergraduate at Changchun Normal University in China, majoring in software engineering. Research interests include deep learning algorithm implementation of intelligent warehouse management system.