Implementation of FaceNet and support vector machine in a real-time web-based timekeeping application

Ly Quang Vu[1], **Phan Thanh Trieu**[2], **Hoang-Sy Nguyen**[3]
[1]Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, Vietnam
[2]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[3]Institute of Engineering and Technology, Thu Dau Mot University, Binh Duong Province, Vietnam

Article Info

Article history:

Received Sep 20, 2021
Revised Dec 20, 2021
Accepted Dec 31, 2021

Keywords:

Face detection
Face recognition
Real-time web-based
Multi-task cascaded neural network
Support vector machine

ABSTRACT

This paper presents in detail how to build up and implement a real-time web-based face recognition application. The system works so that images of people are recorded and compared with the references on the database. If they match, the information about their presence will be recorded. As for the system architecture, the multi-task cascaded neural network was deployed for face detection. Followingly, for the recognizing tasks, we conducted a study to compare the accuracy level of three different face recognizing methods on three different public datasets by means of both the literature review and our simulation. From the comparison, it can be drawn that the FaceNet algorithm in-used with the support vector machine (SVM) classifier performs the best among others and is the most suitable candidate for the practical deployment. Eventually, the proposed system can deliver a highly satisfactory result, proving its potentials not only for the research but also the commercial purposes.

Corresponding Author:

Hoang-Sy Nguyen
Institute of Engineering and Technology, Thu Dau Mot University
Binh Duong Province, Vietnam
Email: nguyenhoangsy@tdmu.edu.vn

1. INTRODUCTION

Face recognition technology has been replacing the human's role in recognizing faces. The face-recognizing equipment receives face images or videos containing human faces as input, of which biometric facial data is subsequently extracted and processed to conduct the recognizing task [1]. Because the facial features of an individual are unique [2], they have been acknowledged as an effective means for security purposes, e.g., alternating the use of passwords and identity cards, and allowing authorized access. Among popular face recognition models which have been developed by universities and companies, there are VGGFace [3], DeepFace [4], [5], OpenFace [6], and FaceNet [7]. In [8], a face recognition model based on the histogram of oriented gradients (HOG) and support vector machine (SVM) classifier was investigated. Besides, in [9], a method based on the AdaBoost algorithm was used to train cascade classifiers with feature types such as the HOG and the Haar-like. Although a better performance was achieved, it is computationally demanding as it includes a number of weak classifiers.

On the other hand, direct training operations on faces can be challenging owing to the face occlusion, which is common in practice. To overcome this issue, Zhang *et al.* [10] have based on the Bayesian framework to propose an algorithm that locates the head using the Omega-liked shape formed by the head-shoulder part of a person. This technique has been applied widely in automatic teller machines (ATM). Additionally, in [11], the face-recognizing task was carried with deformable part models (DPM) yielding remarkable results, though it requires heavy computational resources. The DPM-based system was deployed as well in [12], which offers a reduction in error rate and false-negative face detection. Nonetheless, this technique is limited by the usage of front-view facial images, thus, is not universal.

Recent years have witnessed the rise of convolutional neural network (CNN) application in face detection. Deep CNN (DCNN) [13], region based convolutional neural networks (R-CNN) [14], and another one-or two-stage deep CNN-based systems such as VGGNet [15] and ResNet [16] have showcased their outstanding performance in comparison with their conventional counterparts. However, as there are more convolutional layers added to the CNN, the detecting speed is reduced considerably. To overcome this issue, a number of multi-stage face-detecting algorithms have been investigated, for example, the funnel-structured cascade (FuSt) [17], the pyramid-based cascade model that distills knowledge online and mines hard sample offline [18], which deliver outstanding true positive rate and performance in real-time.

CNNs are driven with data as they are trained with the extracted features and face classification. Additionally, CNNs which are trained with 2D facial data could further be tuned with 3D one for potentially better recognition accuracy. Tornincasa [19] showcased how the pertinent discriminating features from the query faces can be extracted by the use of differential geometry. Dagnes *et al.* [20] have investigated an algorithm that can compute the optimized marker layout to capture the face movement. To deal with the different facial expressions and illumination, radon and wavelet transforms were combined in [21] for the nonlinear feature extraction. Notably, a so-called DeepID model, which is constructed of a large number of CNNs, and its extension were proposed in [22]–[24] with a better feature extracting capability. This is realized thanks to the fact that they can process a variety of face positions and facial patches.

In this paper, we designed a face recognition system based on the FaceNet model with SVM classifier. Then, we compare the accuracy of our proposed method with two other face recognition methods operating on three public datasets to increase the generalization of the study. Finally, the paper showcases how to integrate the system into a web-based timekeeping application. The obtained results regarding the system performance and its implementation are highly applicable for both the research purpose and the practical usage.

2. MATERIALS AND METHODS

2.1. Multi-task cascaded convolutional neural networks (MTCNN)

MTCNN framework detects and aligns faces with unified cascaded CNNs. MTCNN is tasked with three outputs. Firstly, it has to classify whether an input is a face or non-face. Then it has to perform the bounding box regression, and finally localizes the facial landmark. Each layer uses the intakes the output from its preceding layer and in the end, the overall learning target is summed up. Corresponding to these tasks, the MTCNN is constructed of three layers which are in order so-called the proposal network (P-Net), the refine network (R-Net), and the output network (O-Net). The architectures of MTCNN are shown in Figure 1.

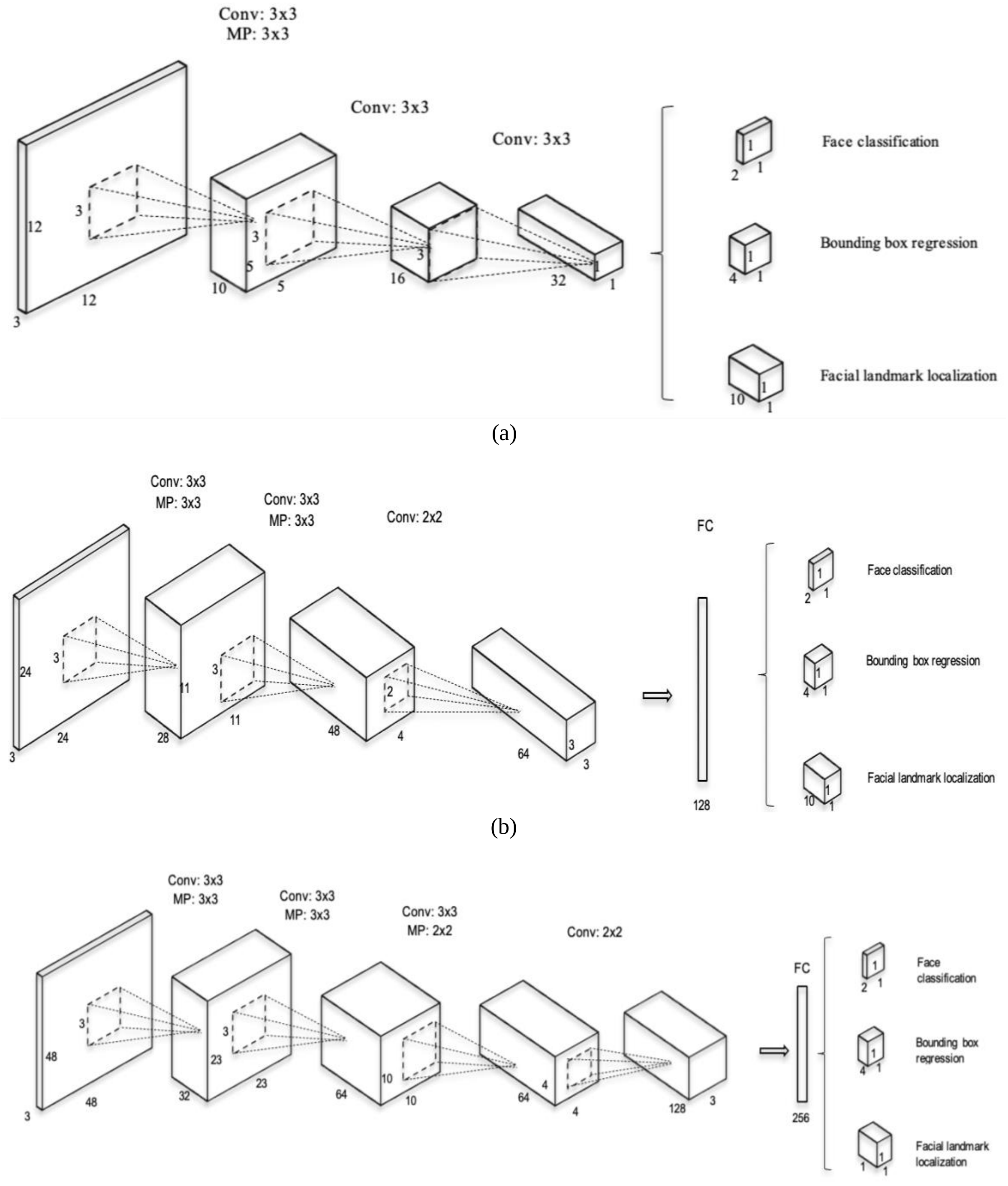
Layer 1(P-Net) is a fully convolutional network (FCN), which is used to generate the candidate windows and their corresponding bounding box regression vectors. P-Net combines the overlapping areas of the bounding box vectors to reduce the candidate volume. Layer 2 (R-Net) is a CNN which is differentiated from FCN as its last stage is denser. R-Net intakes the output of P-Net, screens out the false candidates, calibrates with bounding box regression, and merges overlapping candidates using non-maximum suppression (NMS). Layer 3 (O-Net) functions in a similar manner to the R-Net. However, it describes in more detail the faces and delivers five positions of the facial landmarks being the left/right eyes, nose, and left/right mouth corners.

2.2. FaceNet model

Facenet takes an image of the person's face as input and output it into the 128-dimensional euclidean space. The distances of a person’s face images would be comparatively closer than that of other random ones. In general, there are two different CNN-based basic architectures in Facenet. The first category adds 1×1×d convolutional layers between the standard convolutional layers of the Chen *et al.* [25] architecture, then gets a model 22 layers NN1 model. The second category consists of Inception models based on GoogLeNet [26]. The inception module contains 4 branches from the left to right. It employs convolution with 1×1 filters as well as 3×3 and 5×5 filters and a 3×3 max-pooling layer. Each branch uses a 1×1 convolution to achieve time complexity reduction. FaceNet model is a DCNN trained via a triplet loss technique that allows vectors for the same identity to become more similar, while vectors for different identities should become less similar.

2.3. Triplet loss

The face recognition model is trained with batches of data, each has three images being the anchor, the positive, and the negative images. Specifically, Figure 2 illustrates how the triplet loss operates by maximizing the anchor-negative image distance and minimize the anchor-positive one. Notably, an image is considered positive if it has the same identity as the anchor and vice versa for the negative image. Thanks to this mechanism Triplet loss has been considered one of the best effective ways for learning face image 128-D encodings. Notably, an image is considered positive if it has the same identity as the anchor and vice versa for the negative image. Thanks to this mechanism Triplet loss has been considered one of the best effective ways for learning face image 128-D encodings.



(a)

(b)

(c)

Figure 1. The architectures of (a) P-Net, (b) R-Net, and (c) O-Net, where "MP" means max pooling and "Conv" means convolution. The step size in convolution and pooling is 1 and 2, respectively

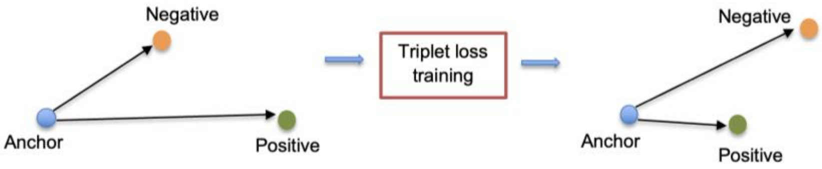


Figure 2. The triplet loss training

2.4. Proposed approach

The pipeline of our face recognition system is illustrated in Figure 3. It can be further elaborated:

- Firstly, the MTCNN is trained with face images of all the staff in an organization.
- Secondly, images and video frames are input into our system and the MTCNN face detector is applied to recognize the face location. These faces are pre-processed and aligned based on the face landmarks computed by MTCNN. There are five features that are included in the face landmarks which are the nose, left/right eye, and left/right mouth. Moreover, the MTCNN is used as well to construct image pyramids corresponding to the face images.
- Thirdly, the FaceNet algorithm is applied to extract the 128D embeddings from the face images.
- Fourthly, we deploy a search algorithm to find in the database an encoding whose distance with the real-time face image encoding satisfies a threshold value. Once it does, the person is recognized as a staff.
- Consequently, the information about the presence of the recognized staff will be updated to the database. Otherwise, the application screens will display a notification announcing that the person's face can not be recognized or it is not in the staff list.

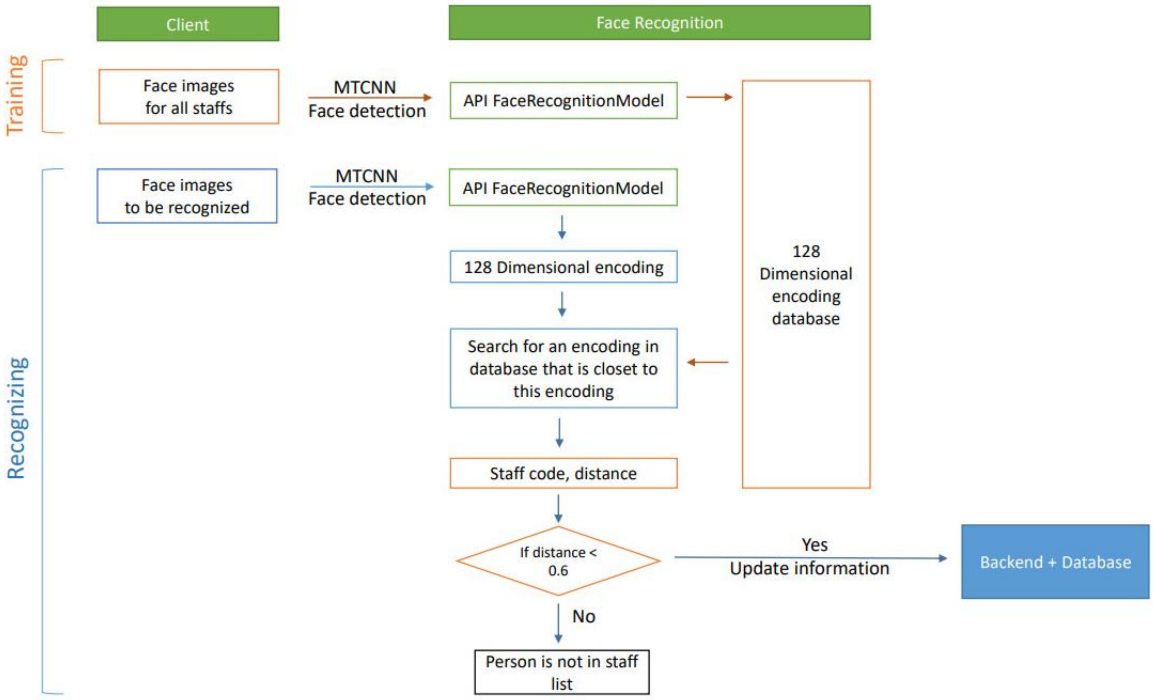


Figure 3. The pipeline of our face recognition system

Remarkably, our system allows working with Euclidean image embeddings, and the network is trained to propose the embedding spaces (squared L2 distances) directly according to the similar faces. As a result, the distance of the same subject images is small and that of different subjects is big. After these embedding spaces have been created, face verification can be performed easily by setting a threshold distance value between two points in the space. Subsequently, the SVM algorithm is applied for the classifying operation.

3. IMPLEMENTATION AND RESULTS

3.1. Datasets used for training and testing

In this paper, three public datasets namely labeled faces in the wild (LFW) [27], our database of faces (ORL) [28], and yale face database [29] were used to assess the accuracy of the in-studied approach. The number of face images and subjects along with their notes are given in Table 1. The three datasets vary largely in the number of images, subjects and configurations. Thus, it is expected that the generality of this study can be ensured.

Table 1. Three in-used public face image datasets

Name	Number of face images and subjects	Challenges	Note
LFW[27]	13,233 images (5,749 subjects)	Face pose, expression, illumination.	Subjects with more than 20 images were selected, resulting in 3,137 images (62 subjects).
ORL[28]	400 images (40 subjects)	Timing, expression (open/close eyes, yes/no smile), illumination, accessories (yes/no glasses).	All subjects were used. Dark background. Upright, frontal face images.
Yale Face Database[29]	165 images (15 subjects)	Expression (happy, neutral, sad, sleepy, surprised, wink), illumination (center/left/right light), accessories (yes/no glasses).	All subjects were used. Grayscale GIF images.

3.2. Experiment results

Table 2 compares the accuracy of the results which were produced from the three datasets using the FaceNet with support vector machine (SVM) classifier. In particular, every image was processed with a Euclidean space technique and compared with its index label. It can be concluded that the FaceNet with SVM can deliver results with relatively a high level of accuracy. Figure 4 demonstrates how the triplet loss function minimize the distances between positive anchors and maximizes the distances between negative ones after being trained with the subset of the LFW dataset.

Table 2. Accuracy comparison using Facenet with SVM

Dataset	FaceNet with SVM [%]
LFW	99.83
ORL	97.5 [2]
Yale Face Database	98.9 [2]

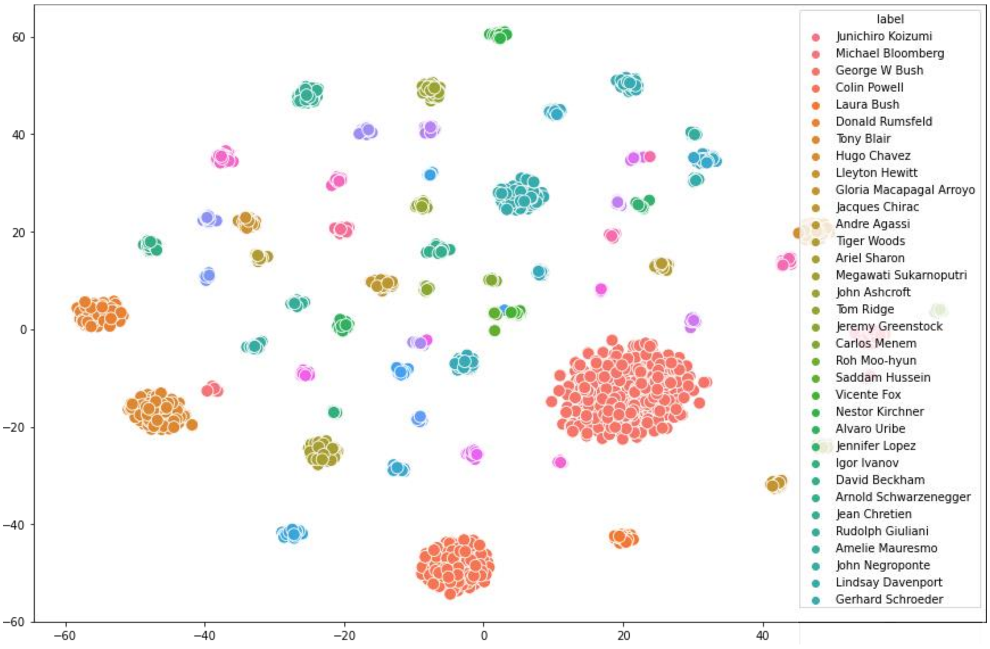


Figure 4. Triplet loss training on the subset of LFW dataset

The results from the FaceNet face recognition is subsequently compared with two other methods namely the principal component analysis (PCA) and SVM classifier [25], and the k-nearest neighbor (K-NN) [26], as can be seen in Table 3, the FaceNet with SourVM can deliver the minimum accuracy level of up to 97.5%, being the highest among others. It should be noted that this model performs well even though there exist challenges being a variety of face poses, expressions, illumination, and the use of accessories.

Table 3. Accuracy comparison using FaceNet with SVM, PCA with SVM, and K-NN

Method Dataset	LFW [%]	ORL [%]	Yale Face Database[%]
FaceNet with SVM	99.83	97.5 [2]	98.9 [2]
PCA with SVM	62.14	95.12	82.35
K-NN	30.24	85.36	52.94

Face recognition can effectively detect human presence in a particular area of interest (AOI) such as office, and educational institution. Herein this paper, the authors succeeded in establishing a web-based timekeeping application. Figure 5 illustrates how the system works. The system consists of a remote server and a database that can be accessed with a web application for monitoring and administrating purposes. An IP camera is set at the entrance to a company to streamline video frames in real-time to the Face recognition API. If a face is detected, an image in that time frame is preprocessed and passed on to the deep CNN to generate 128-byte embedding.

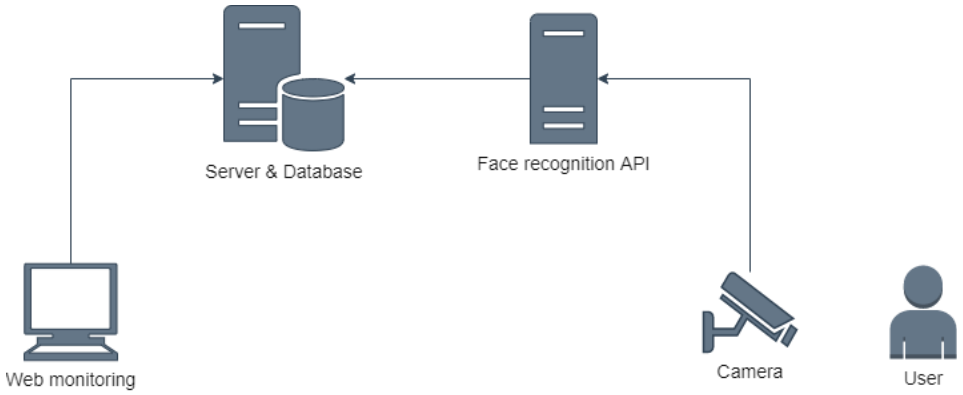


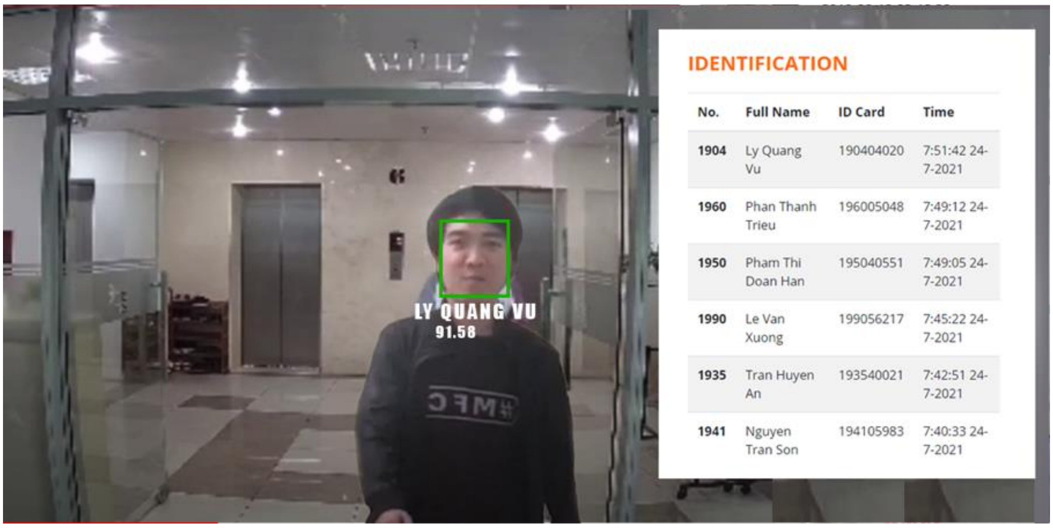
Figure 5. Web-based timekeeping application

Subsequently, the staff's identities can be determined with the SVM classifier and the data related to the staff's presence such as the identities, the accuracy percentage, the time, and the date of entry are recorded in the database. Figure 6 illustrates what information a user can see on the web application as a staff is recognized by the system. Specifically, there is a frame identifying the detected face with the recognized name and the accuracy at its bottom. The right side shows a list of recognized staffs along with their ID numbers, full names, ID cards, and the entry time. In case the system cannot recognize a person's presence due to the missing of data, for example, an entry of new staff or a visitor, the face image of the person will display as shown in Figure 7.

The detected face is framed with red color and labeled with "Unknown". It should be noted that the time and date of the unrecognized entry are recorded to assist the administrator in preparing corresponding solutions such as adding the information of the new employee, and re-training the model. All the information about the entries of the people as in Figure 6 can be exported to *.xls file as shown in Figure 8.

Besides, the web application has an interface for adding new facial data. Users can open the IP camera from the application to capture new face images in real-time. These images can then be saved in the database and assigned with a unique user ID. Consequently, the face from the image is extracted and labeled with what the administrator may find suitable, for example, the person's name.

The system was tested with a group of 32 staffs showing the face recognition accuracy of 96%. Nevertheless, the system is sensitive to the lighting changes and angle between the faces and the IP camera, which considerably downgrade the system's accuracy. Thus, in case the system fails to recognize staff, the staff needs to inform the person in charge of timekeeping for a manual marking.



IDENTIFICATION

No.	Full Name	ID Card	Time
1904	Ly Quang Vu	190404020	7:51:42 24-7-2021
1960	Phan Thanh Trieu	196005048	7:49:12 24-7-2021
1950	Pham Thi Doan Han	195040551	7:49:05 24-7-2021
1990	Le Van Xuong	199056217	7:45:22 24-7-2021
1935	Tran Huyen An	193540021	7:42:51 24-7-2021
1941	Nguyen Tran Son	194105983	7:40:33 24-7-2021

Figure 6. A recognized staff by the system

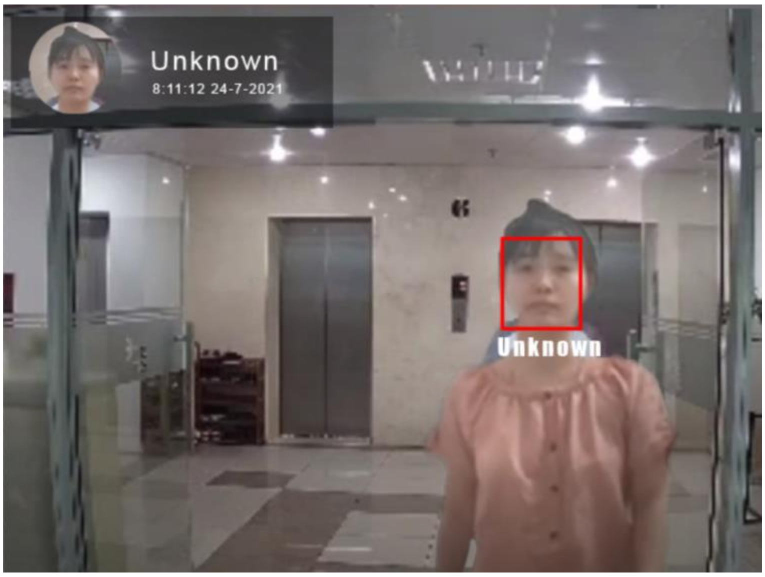
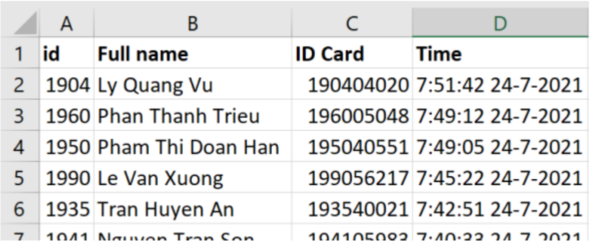


Figure 7. An unrecognized person



	A	B	C	D
1	id	Full name	ID Card	Time
2	1904	Ly Quang Vu	190404020	7:51:42 24-7-2021
3	1960	Phan Thanh Trieu	196005048	7:49:12 24-7-2021
4	1950	Pham Thi Doan Han	195040551	7:49:05 24-7-2021
5	1990	Le Van Xuong	199056217	7:45:22 24-7-2021
6	1935	Tran Huyen An	193540021	7:42:51 24-7-2021
7	1941	Nguyen Tran Son	194105983	7:40:33 24-7-2021

Figure 8. Data exported to *.xls file.

4. CONCLUSION

To conclude, in our system, the MTCNN algorithm is deployed to detect the faces, generates the embeddings using the pre-trained FaceNet with SVM classifier, then recognizes images that are taken through the system. The system is able to deliver in practice the recognition accuracy of 96%, given that the images are collected under consistent conditions in terms of lighting and face-camera angle. The comparison study can serve as a foundation for the researchers seeking for optimized face-recognizing algorithms. Additionally, the paper also presents an established web-based application with some key concepts that can potentially be upgraded to a commercial timekeeping product. Application of such products into practice has proven its abilities to save companies and organizations a considerable amount and time and efforts in timekeeping tasks. As more and more powerful algorithms are introduced and implemented into face recognition systems, it is promising that end users will get more benefits from them. For future studies, the system can be more fine-tuned and more training data with noises can be collected to further improve the capability of our proposal.

REFERENCES

[1] Y. Zhang, S. Wang, H. Xia, and J. Ge, "A novel SVPWM modulation scheme," in *2009 Twenty-Fourth Annual IEEE Applied Power Electronics Conference and Exposition*, Feb. 2009, pp. 128–131, doi: 10.1109/APEC.2009.4802644.

[2] L. Li, X. Mu, S. Li, and H. Peng, "A review of face recognition technology," *IEEE Access*, vol. 8, pp. 139110–139120, 2020, doi: 10.1109/ACCESS.2020.3011028.

[3] I. William, D. R. Ignatius Moses Setiadi, E. H. Rachmawanto, H. A. Santoso, and C. A. Sari, "Face recognition using facenet (survey, performance test, and comparison)," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICIC47613.2019.8985786.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Procedings of the British Machine Vision Conference 2015*, 2015, pp. 41.1-41.12, doi: 10.5244/C.29.41.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1701–1708, doi: 10.1109/CVPR.2014.220.

[6] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10, doi: 10.1109/WACV.2016.7477553.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[8] M. Drożdż and T. Kryjak, "FPGA implementation of multi-scale face detection using HOG features and SVM classifier," *Image Processing & Communications*, vol. 21, no. 3, pp. 27–44, Sep. 2016, doi: 10.1515/ipc-2016-0014.

[9] C. Ma, N. Trung, H. Uchiyama, H. Nagahara, A. Shimada, and R. Taniguchi, "Adapting local features for face detection in thermal image," *Sensors*, vol. 17, no. 12, Art. no. 2741, Nov. 2017, doi: 10.3390/s17122741.

[10] T. Zhang, J. Li, W. Jia, J. Sun, and H. Yang, "Fast and robust occluded face detection in ATM surveillance," *Pattern Recognition Letters*, vol. 107, pp. 33–40, May 2018, doi: 10.1016/j.patrec.2017.09.011.

[11] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Computer Vision {\textendash} {ECCV} 2014*, Springer International Publishing, 2014, pp. 720–735.

[12] D. Marcetic and S. Ribaric, "Deformable part-based robust face detection under occlusion by using face decomposition into face components," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2016, pp. 1365–1370, doi: 10.1109/MIPRO.2016.7522352.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[14] S. Wan, Z. Chen, T. Zhang, B. Zhang, and K. Wong, "Bootstrapping face detection with hard negative examples," Aug. 2016.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, Available: http://arxiv.org/abs/1409.1556.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[17] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascade for multi-view face detection with alignment-awareness," *Neurocomputing*, vol. 221, pp. 138–145, Jan. 2017, doi: 10.1016/j.neucom.2016.09.072.

[18] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, Jun. 2015, pp. 643–650, doi: 10.1145/2671188.2749408.

[19] S. Tornincasa *et al.*, "3D facial action units and expression recognition using a crisp logic," *Computer-Aided Design and Applications*, vol. 16, no. 2, pp. 256–268, Aug. 2018, doi: 10.14733/cadaps.2019.256-268.

[20] N. Dagnes *et al.*, "Optimal marker set assessment for motion capture of 3D mimic facial movements," *Journal of Biomechanics*, vol. 93, pp. 86–93, Aug. 2019, doi: 10.1016/j.jbiomech.2019.06.012.

[21] H. D. Vankayalapati and K. Kyamakya, "Nonlinear feature extraction approaches with application to face recognition over large databases," in *2009 2nd International Workshop on Nonlinear Dynamics and Synchronization*, Jul. 2009, pp. 44–48, doi: 10.1109/INDS.2009.5227967.

[22] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: face recognition with very deep neural networks," Feb. 2015, [Online]. Available: http://arxiv.org/abs/1502.00873.

[23] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," Jun. 2014, [Online]. Available: http://arxiv.org/abs/1406.4773.

[24] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1891–1898, doi: 10.1109/CVPR.2014.244.

[25] X. Chen, L. Song, and C. Qiu, "Face recognition by feature extraction and classification," in *2018 12th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, Nov. 2018, pp. 43–46, doi: 10.1109/ICASID.2018.8693198.

[26] H. Zhang and G. Chen, "The research of face recognition based on PCA and k-nearest neighbor," in *2012 Symposium on Photonics and Optoelectronics*, May 2012, pp. 1–4, doi: 10.1109/SOPO.2012.6270975.

[27] B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," 2008.

[28] "ORL (our database of faces)." https://paperswithcode.com/dataset/orl.

[29] "Yale face database." http://vision.ucsd.edu/content/yale-face-database.

BIOGRAPHIES OF AUTHORS

Ly Quang Vu is a student Master of Science (M.Sc.) in the Faculty of Computer Science of Ho Chi Minh City University of Information Technology (VNUHCM-UIT). He is working at Bdoop Services and Trading Joint Stock Company-Ho Chi Minh City, Vietnam. His research interests are in fields of Machine Learning Applications and Image Processing, Data Mining, and Network Security. Email: quangvu.ly@gmail.com.

Phan Thanh Trieu is a student Master of Science (M.Sc.) in the Faculty of Information Technology (IT) of Ton Duc Thang University, Vietnam (TDTU). He is working at Vietnam Posts and Telecommunications Group (VNPT)-An Giang Province, Vietnam. His research interests are in fields of Machine Learning Applications and Image Processing, Network Communications, and Network Security. Email: phanthanhtrieuag@gmail.com.

Hoang-Sy Nguyen was born in Binh Duong province, Vietnam. He received the B.S. and MS.c degree from the Department of Computer Science from Ho Chi Minh City University of Information Technology (UIT-HCMC), Vietnam in 2007, 2013, respectively. He received his Ph.D. degree in communication technology, dissertation thesis "Energy harvesting enable relaying networks: Design and performance analysis" from the VSB-Technical University of Ostrava-Czech Republic, in 2019. His research interests include Energy efficient wireless communications, 5G wireless communication networks, Network security, Artificial Intelligence, Cloud Networks, and Big Data. Email: ng.hoangsy@gmail.com.