Vol. 04, No. 4,(2020),162–170, ***ISSN:2581-3242***

IJMLNCE
JOURNAL

International Journal of Machine Learning and Networked Collaborative Engineering

Journal Homepage: http://www.mlnce.net/home/index.html

DOI : https://doi.org/10.30991/IJMLNCE.2020v04i04.003

Improving the Performance of One-shot Face Recognition by using Data Augmentation

[a]Nguyen Thanh Trong, [b]Luong Gia Kien, [c]Thi T. T. Tran, [d]Hieu N. Duong, [e]Tran Van Hoa, and [f]Thoai Nam

[a]Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam,1870088@hcmut.edu.vn

[b]Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, giakienlgk@gmail.com

[c]Hoa Sen University, Ho Chi Minh City, Vietnam ,thi.tranthitruong@hoasen.edu.vn

[d]Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, dnhieu@hcmut.edu.vn

[e]Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, tranvanhoa533@gmail.com

[f]Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, namthoai@hcmut.edu.vn

Abstract

For a past few years, the revolution of deep learning techniques has emerged and launched several state-of-the-art models, for instance, the breakthroughs of DeepFace and DeepID to face recognition in 2014. The face recognition in CCTV systems commonly encounters a few obstacles coming from practical conditions, such as ambient light, the diverse positions and angles of cameras, face masks, face poses, and so on. In addition, people who are monitored by the CCTV systems lack photos and typically have only one photo. These problems lead to face recognition reported with unstable performance and difficult to be successfully used in practice. To tackle these problems, this paper proposes an approach, namely ISE, to face augmentation which interpolates multiple samples from an original photo. Particularly, the samples produced by ISE contain real characteristics of cameras in the CCTV systems. By practically deploying a CCTV system at the Bach Khoa Dormitory, ISE indicates that it can boost the performance of face recognition up from 72%, 46% to 84%, 64% in daytime and day-and-nighttime, respectively.

Keywords

One-shot face recognition,

CCTV,

ISE,

ASE.

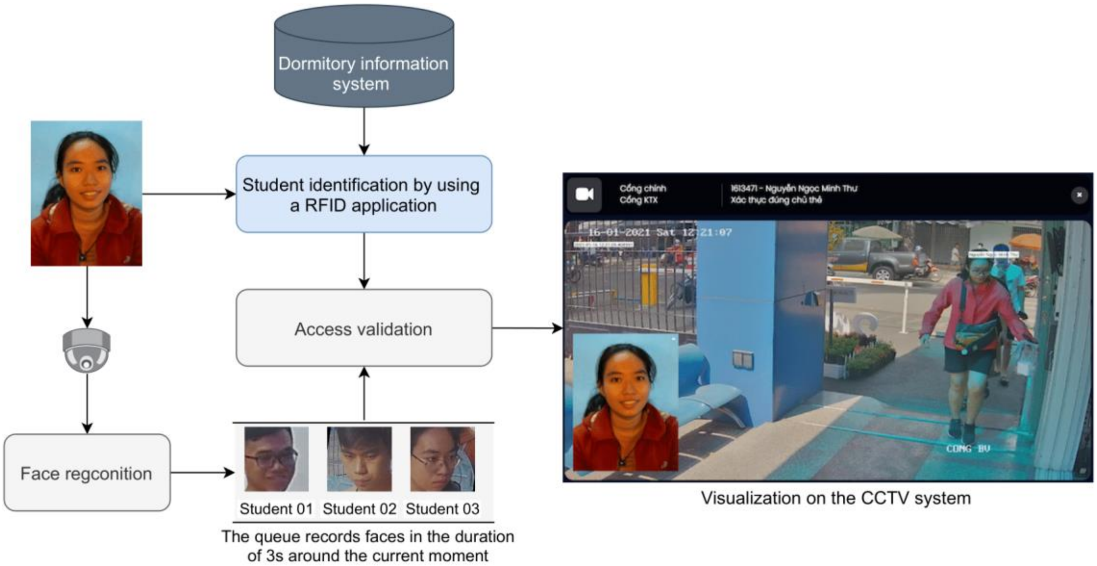
1. Introduction

Recently, the use of facial recognition systems to address security problems has been considered by several scientists, organizations and governments in the world. These problems are diverse, such as criminal detection, access control systems, electronic passport gate, and so on. Due to the innovation of deep-learning techniques and powerful hardware, multiple computer vision tasks become easy to be tackled and the concept of one-shot learning for face recognition has been proposed recently [1–6].

Generally, one-shot learning is a classification task where one, or a few examples are used to classify many new examples in the future. For the tasks of face recognition, a person can be classified correctly with one or a few photos. It is significantly challenging since the performance of face recognition models depend on several factors, such as face pose, light intensity, expression, and so on. So that one or a few photos do not provide enough information for the models to recognize the faces, especially to identify the faces against a specific database. Unfortunately, most practical databases just store a photo of an individual, such as the human resource database of a company, criminal databases, and so on. It leads to a great deal of challenges since the face recognition combined with Closed-circuit television (CCTV) systems, for instance, how to scan crowds captured by the CCTV systems and identify correctly faces in these crowds against a database of known or suspected criminals.

To tackle the problem of lacking sample photos, data augmentation techniques are often used to enrich the information of a given template photo. There are several data augmentation techniques that are classified into three groups consisting of generic transformation, component transformation, attribute transformation [7]. However, these techniques are more relevant to enriching datasets used in training phases than to inferencing phases. This paper introduces an approach which reinforces new samples of individuals. The reinforcement process will produce multiple new samples based on the difference between avatar faces and the sample faces captured by the CCTV systems.

The problem arises from a CCTV system that has been deployed at the Bach Khoa Dormitory – the dormitory of Ho Chi Minh University of Technology, Ho Chi Minh City, Vietnam. The CCTV system tackles three security tasks: 1) access verification; 2) suspected criminal monitoring; and 3) restricted area monitoring. For the first task, when students cross the main entrance of the dormitory, the students must use student cards which are RFID cards to verify who they are. Typically, security guards working at the entrance validate if the actual students and the faces shown on a RFID software are the same ones. Consequently, it is challenging for the security guards to handle all faces of people simultaneously passing the entrance. Hence, the system integrated face recognition engine plays the role of security guards and automatically checks if ones are students staying at the dormitory when they cross the entrance. In the RFID software database, each student just has a photo as their avatar. At the entrance, two cameras were equipped to capture all people passing the entrance. The figure 1 presents the workflow of access validation since one crosses the entrance.


Dormitory information system
Student identification by using a RFID application
Access validation
Face regconition
Student 01
Student 02
Student 03
The queue records faces in the duration of 3s around the current moment
Visualization on the CCTV system

Cổng chính
Cổng KTX
1613471 - Nguyễn Ngọc Minh Thư
Xác thực đúng chủ thể
16-01-2021 Sat 12:21:07

Figure 1: The workflow to validating accesses of students at the main entrance.

Aiming for comfortability to people who cross the entrance, the identification task attempts in the best way to reduce interactions between students and the CCTV system as much as possible. It means people do not realize that they are monitored by cameras. When they cross the entrance, they just give their RFID cards into RFID card readers and pass if the card is valid. During crossing, their faces captured by cameras have arbitrary features, such as pose, light, taking accessory, and so on. It leads to three main issues as follows.

- First of all, the accuracy of face recognition to students having just an avatar is often lower than expectation. As a practical statistic during a few deploying months, the true acceptance rate of the face identification is approximately $50\% - 70\%$. The main reason is that facial characteristics of the sample photos (avatar) and real photos captured by the CCTV system are actually different due to face pose, light condition, changes of faces by time, and so on.
- The second, to tackle the first challenge, the system is designed to automatically collect actual face photos of all students staying at the dormitory. During students cross the entrance, if the

CCTV system captures and correctly identifies students against the database, multiple their face photos will be captured and stored as new sample photos. By time, the sample photos of students are increased. Experiments indicate that if an individual has one hundred sample photos, the true acceptance rate is improved approximately 68% − 84%. However, the process of photo collection takes a few weeks to finish collecting one hundred samples for a student.

- The third, the fluctuation of students at the dormitory is large since an approximate amount of 40 percent students staying at the dormitory annually graduate and leave. Simultaneously, the dormitory receives the same amount of new students. As a result, it takes a few weeks to collect actual samples for the new students. During this period, the system often incorrectly identifies new students due to the lack of sample photos.

In this paper, we introduce a method called Interpolated Sample Enhancement (ISE), which generates augmented samples of individuals from an avatar. The performance of ISE will be practically measured at the Bach Khoa Dormitory. The rest of this paper is organized as follows. Section 2 presents two methods to face augmentation, especially explains in detail the proposed approach based on clustering algorithms. In section 3, we present experimental results. Some findings also will be indicated in Section 3. Finally, we draw conclusion and perspectives for future work in Section 4.

2. Methodolody

2.1 Two methods of face augmentation

As presented in the introduction part, the most challenging task of the CCTV system deployed at Bach Khoa Dormitory is how to construct new samples for new students from an avatar photo. These problems are tackled by two methods namely ASE and ISE. ASE and ISE do not independently work but interwoven with each other. ASE samples provide information to build ISE samples and then the ISE samples are used to build the ASE samples of new students.

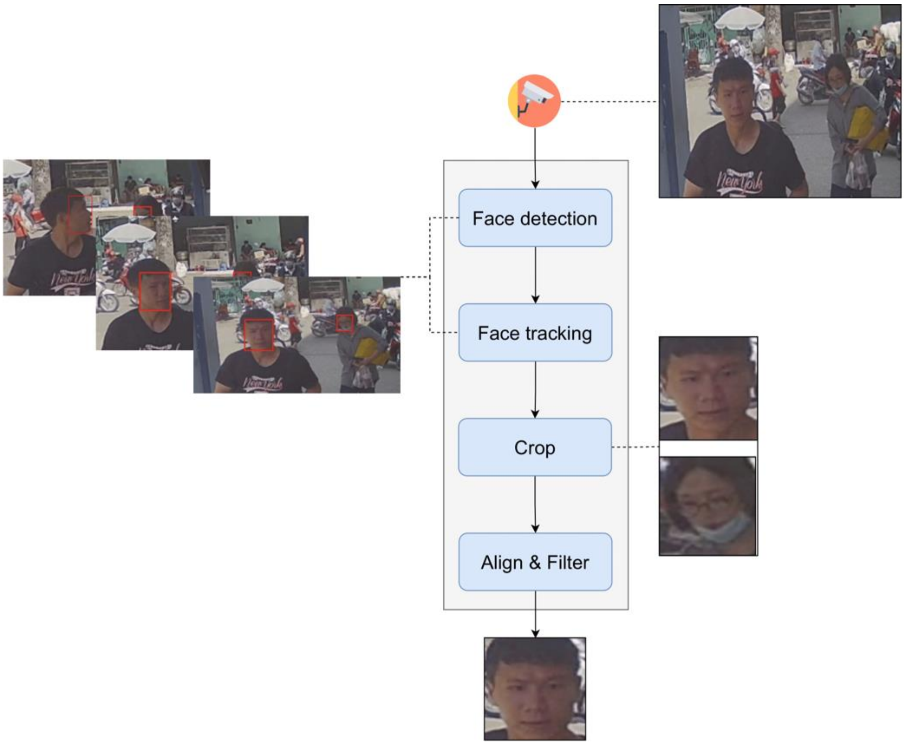


Figure 2: Flow of face detection.

- **Actual Sample Enhancement (ASE) method**. This method is applied every time the known students appear in front of the CCTV system. While students appear on cameras, their faces are captured and tracked by multiple continuous frames until they vanish from the cameras. The Retinaface model and the Kalman filter algorithm are employed to detect and track faces on cameras, respectively [8,9]. In these frames, the faces may be different from poses and sizes. A few heuristic algorithms based on facial landmarks and image processing techniques are used to filter all faces which are not good at poses, sizes and light. Figure 2 presents the flow of the face detection process. Then, the remaining faces are encoded to embedding vectors and to be searched against the student database. The face encoder is a deeplearning model called ArcFace [10]. Cosine measure is used to identify the similarity of two embedding vectors. If one recorded sample of a student is similar to the embedding vectors, the embedding vectors are added into the database as new samples of this student, namely ASE samples. By employing the ASE process, new students in the system can easily be added enough samples when they frequently appear in front of the CCTV system. The diverse characteristics of ASE samples can improve the true acceptance rate of identification tasks. The maximum number of ASE samples recored to the database is 100. The number is sufficient to balance between the searching time and accuracy of face recognition. Figure 2 illustrates ASE samples of an avatar. Note that there are approximately 3000 students staying at the dormitory. Thus the number of all samples in the database reaches to 300, 000. To accelerate searching, we employ Vearch[*] which is based on KGraph [11] to store all the samples.

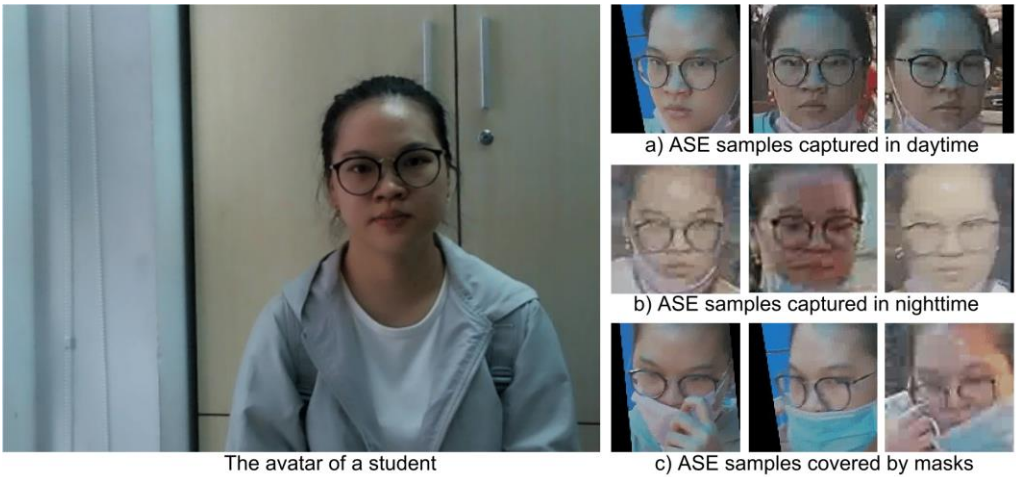

a) ASE samples captured in daytime
b) ASE samples captured in nighttime
The avatar of a student
c) ASE samples covered by masks

Figure 3: An illustration of actual samples of a student gathered by ASE.

- **Interpolated Sample Enhancement (ISE) method.** The major principle of ISE is based on the difference between the distributions of avatar photos and actual face photos captured by the CCTV system. The correct extraction of the difference is crucial to the effectiveness of ISA. The difference is caused by multiple factors, such as poses, light, change by time of face, and so on. For instance, Figure 2 depicts the diversity of actual faces captured by the CCTV system. ISE utilizes a famous clustering algorithm namely K-means [12] to find out the difference and produces new samples by adding the difference to the avatars.

2.2 ISE

The main idea of ISE is summarized as in Figure 4. Assume that actual and avatar embedding vectors are distributed in several difference spaces. The larger differences are, the lower performance of face recognition is. Thus, ISE method attempts to figure out the differences and uses the differences to infer how actual faces of an avatar are. In other words, ISE samples are the projection of avatars from avatar distributions to actual face distributions. The process of ISE consists of three steps as follows.

- *Step 1 − calculating delta vectors.* Each avatar has a corresponding embedding vector which was encoded by the ArcFace model. This avatar has a set of ASE samples gathered by the CCTV

[*] https://github.com/vearch/vearch

system. These ASE samples are also encoded to ASE embedding vectors. Delta vectors of an avatar are obtained from the difference of the avatar embedding vector and the ASE embedding vectors of this avatar. Equation1 briefly describes how to calculate all delta vectors.

$$\text{delta\_vectors} = \text{avatar\_embedding\_vectors} - \text{ASE\_embedding\_vectors} \qquad (1)$$

- *Step 2 − clustering the delta vectors.* The K-means algorithm is used to cluster all the delta vectors to produce K clusters whose centers present for the mentioned difference. Euclidean distance is used to measure distances of delta vectors during executing the K-means algorithm.
- *Step 3 − constructing ISE sample embedding vectors.* The new sample embedding vectors are constructed by plussing embedding vectors of avatars and vectors of cluster centers as in Equation2. To avoid noise, the new embedding vectors are validated by comparing with the embedding vectors of avatars. All unsimilar new embedding vectors to avatars are removed.

$$\text{ISE\_embedding\_vectors} = \text{avatar\_embedding\_vectors} + \text{delta\_vectors} \qquad (2)$$

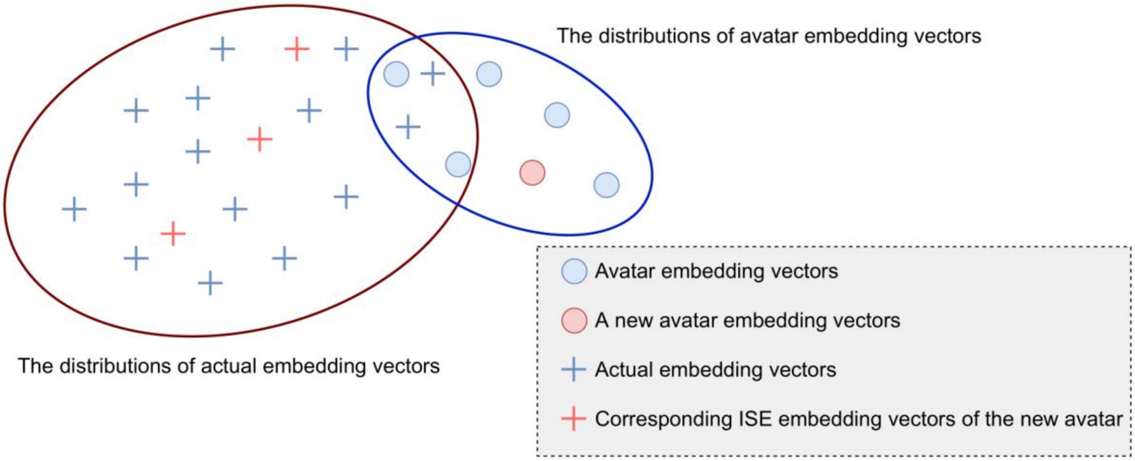

The distributions of avatar embedding vectors
The distributions of actual embedding vectors
○ Avatar embedding vectors
○ A new avatar embedding vectors
+ Actual embedding vectors
+ Corresponding ISE embedding vectors of the new avatar

Figure 4: The conceptual model of ISE.

In practical deployment, we organize new samples into two spaces to store samples collected by ASE and ISE. ASE samples are real and better than ISE samples in terms of accuracy. Hence, ISE samples are just used to search for new students lacking ASE samples. If a student has the amount of ASE samples larger than a threshold (e.g., 20), the identification of this student is only executed on ASE samples. It means that ISE is a temporarily alternate of ASE in short term.

3. Experiments

The dataset used for constructing ISE model was collected within about one month. The dataset contains 32401 actual face photos of 1561 students after about 10670 access. In average, each student is recorded about three new samples for one access, and has about 20 actual samples gathered by the ASE method alongside with his/her avatar. We use the dataset for two purposes that are i) to build the clusters used to construct ISE samples and ii) to figure out the most appropriate cosine threshold used to decide if two embedding vectors are similar.

- Building clusters by using the K-mean algorithm. The drawback of K-mean is to identify the optimal K value. Typically, the larger K value is, the larger both TAR and FAR are. In other words, K value is large, the amount of ISE samples is large too. It leads to the information of ISE samples becoming more diverse but noisier too. Otherwise, K value is small, the information of ISE samples is insufficient. Hence the K value of 50 is set to balance TAR and FAR, in practice at the Bach Khoa dormitory.

- Figuring out the approximate optimal cosine threshold in terms of analyzing the performance of face identification with only an avatar. Figure 5 and table 1 indicate that increasing the cosine threshold leads to both TAR and FAR increasing too. To balance the values of TAR and FAR, the optimal cosine threshold of 0.5 is chosen since the values of TAR and FAR are 52.57% and 0.43%, respectively.

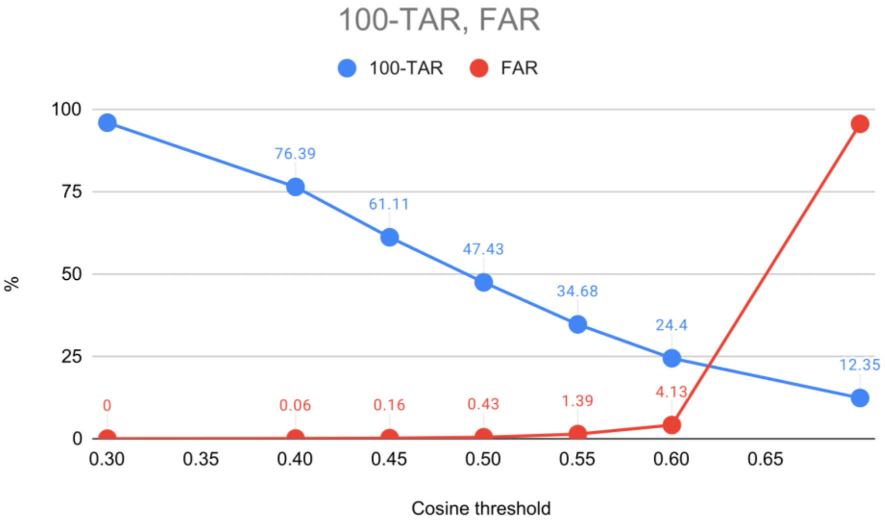


Figure 5: The illustration of correlation between cosine threshold and TAR, FAR.

ASE and ISE are integrated into a dedicated CCTV system and practically tested at the Bach Khoa dormitory. We set up three scenarios to test the performance of ASE and ISE. The CCTV system was set to run within two weeks. During this duration, the system counts students who cross the main entrance and have i) only an avatar in the database; ii) an avatar and 100 ASE samples and iii) an avatar and ISE samples. For each student access, the system recorded the status of recognition consisting of true acceptance and false rejection. Then these numbers are used to independently measure the performance of the system without ASE and ISE, aided by ASE and by ISE. Experimental metrics are divided into two durations of daytime and nighttime to clarify how the light factor impacts the performance of face recognition. Table 2 presents the results of practical tests in duration of two weeks and in daytime – between 7AM and 5PM. In the daytime, the sunlight is good enough to support cameras working well. Meanwhile in nighttime, motion capturing of cameras is not good and cameras often capture blurred faces as illustrated in Figure 3 c). It leads to the performance of face recognition unstable and much different between daytime and nighttime. Table 2 describes the performance of face recognition in duration of two weeks including nighttime.

The results in Table 2 and Table 3 also indicate that the performance of the system integrated by ASE and ISE are nearly equal and outperform to none of ASE and ISE integration. Specially, ISE boosts the performance of face recognition up from 71.86%, 46.70% to 83.46%, 63.38% in daytime and day-and-nighttime, respectively. Besides, the results also show that the performance of ASE integration is somewhat better than ISE integration, especially in nighttime.

Table 1: The correlation between cosine threshold and TAR, FAR to 10670 accesses.

Cosine threshold	True acceptance	TAR	False acceptance	FAR
0.3	440	4.12	0	0
0.4	2519	23.61	6	0.06
0.45	4150	38.89	17	0.16
0.5	5609	52.57	46	0.43
0.55	6970	65.32	148	1.39
0.6	8067	75.6	441	4.13

0.7	9352	87.65	10192	95.52

For security tasks involving access control, FAR is one of important indexes that needs to be regarded. False Acceptance Rate (FAR) is the percentage of identification in which unauthorized persons are incorrectly accepted. Although ISE and ASE significantly improve the performance of one-shot face recognition, if FAR index is large, ISE and ASE are inapplicable. To measure FARs of ISE and ASE, we decide to manually assess case by case at the main entrance for a few days. A group of students are intentionally removed from the database of CCTV system and play the role of strangers. When these students access the entrance, we check if the CCTV system misidentifies these students as any other students who reinforced by ASE or ISE samples and are in the database. The results indicate that both FARs of ASE and ISE are less than 1%. Hence, ASE and ISE are applicable to dedicated CCTV systems.

Table 2: Results of practical tests in the duration of two weeks and in daytime between 7AM and 5PM.

Cosine threshold	Access	True acceptance	False rejection	TAR
No ASE & ISE	4186	3008	1178	71.86%
ASE	5924	4991	933	84.25%
ISE	3591	2997	594	83.46%

Table 3: Results of practical tests in the durations of two weeks including night-time.

Cosine threshold	Access	True acceptance	False rejection	TAR
No ASE & ISE	8564	3999	4565	46.70%
ASE	13670	9347	4323	68.38%
ISE	5029	3202	1827	63.67%

4. Conclusion

In this paper we introduced an approach of data augmentation to face recognition namely ISE. ISE is based on the transformation from avatar space to actual face space to address the problems of one-shot face recognition. Although ISE is uncomplicated and understandable, the effectiveness of ISE is significantly regarded. The major drawback of ISE is that ISE should be rebuild for every new CCTV system to fully understand the new distributions of faces captured by this CCTV system. In other words, to build ISE samples we need new ASE samples to provide new information of the new CCTV system and it takes time to gather enough ASE samples. However, this problem can be addressed by time since many CCTV systems are deployed and ASE samples become big and diverse. For future work, how to produce ISE samples based on tiny datasets of ASE samples is also taken into consideration. Besides we consider to equip cool-white lamps and modern cameras at the entrance to improve the performance of the system in nighttime.

ACKNOWLEDGMENT

This work is funded by the Department of Science and Technology, Ho Chi Minh City, Vietnam, under Contract No.127/2019/HD-QPTKHCN issued on December 27, 2019.

References

[1] Yaniv Taigman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1701–1708, 2014.

[2] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification–

verification. arXiv preprint arXiv:1406.4773, 2014.

[3] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint arXiv:1506.07310, 2015.

[4] Matthew Turk and Alex Pentland. Eigenfaces for recognition. Journal of cognitive neuroscience, 3(1):71–86, 1991.

[5] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3025–3032, 2013.

[6] Mei Wang and Weihong Deng. Deep face recognition with clustering based domain adaptation. Neurocomputing, 393:1–14, 2020.

[7] Xiang Wang, Kai Wang, and Shiguo Lian. A survey on face data augmentation for the training of deep neural networks. Neural computing and applications, pages 1–29, 2020.

[8] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641, 2019.

[9] Zaheer Shaik and Vijayan Asari. A robust method for multiple face tracking using kalman filter. In 36th Applied Imagery Pattern Recognition Workshop (aipr 2007), pages 125–130. IEEE, 2007.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4690–4699, 2019.

[11] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th international conference on World wide web, pages 577–586, 2011.

[12] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA, 1967.

Author’s Biography

Nguyen Thanh Trong is AI Engineer of the Center of Computer Engineering, Ho Chi Minh City University of Technology. He received his bachelor from Ho Chi Minh City University of Technology in 2017 and have studied M.S at Ho Chi Minh University of Technology, Ho Chi Minh City, Vietnam since 2018. His research interests are data science, computer vision.

Email: 1870088@hcmut.edu.vn

Luong Gia Kien is AI Engineer of the Center of Computer Engineering, Ho Chi Minh City University of Technology. He received his bachelor from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam in 2018. His research interests are data science, computer vision.

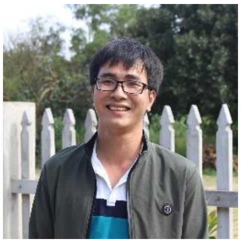
Email: giakienlgk@gmail.com

Tran Thi Truong Thi is lecturer of Hoa Sen University, Ho Chi Minh City, Vietnam. She received her M.S from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam in 2005. Her research interests are data science, visual question answering (VQA).

Email: thi.tranthitruong@hoasen.edu.vn

Hieu N. Duong is deputy director of the Center of Computer Engineering, Ho Chi Minh City University of Technology. Before playing the role of deputy director, he used to be a lecturer at the Falculty of Computer Science and Engineering, Ho Chi Minh City University of Technology from 2002-2018. He received his PhD from VSB-Technical University of Ostrava, Czech Republic in 2016. His research interests are Internet of Things, data science.

Email: dnhieu@hcmut.edu.vn



Tran Van Hoa is AI Engineer of the Center of Computer Engineering, Ho Chi Minh City University of Technology. He received his M.S from Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam in 2020. His research interests are data science, computer vision.

Email: tranvanhoa533@gmail.com

Thoai Nam is director of the Center of Computer Engineering, Ho Chi Minh City University of Technology. Before playing the role of director, he used to be dean of the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology from 2007-2018. He received his PhD from Johannes Kepler University of Linz – Austrian republic.

Email: namthoai@hmcut.edu.vn