Realtime face matching and gender prediction based on deep learning

Thongchai Surinwarangkoon[1], Vinh Truong Hoang[2], Ali Vafaei-Zadeh[3], Hayder Ibrahim Hendi[4], Kittikhun Meethongjan[5]

[1]Department of Business Computer, College of Innovation and Management, Suan Sunandha Rajabhat University, Bangkok, Thailand
[2]Department of Computer Vision, Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[3]Graduate School of Business, Universiti Sains Malaysia, Penang, Malaysia
[4]Department of Information Technology, Computer and Math College, University of Thi Qar, Thi Qar, Iraq
[5]Department of Computer Science, Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand

Article Info

Article history:

Received Aug 4, 2022
Revised Dec 3, 2022
Accepted Dec 7, 2022

Keywords:

Face analysis
Face recognition
Gender prediction
Image classification
Information prediction
Multitask cascade
Convolutional neural network

ABSTRACT

Face analysis is an essential topic in computer vision that dealing with human faces for recognition or prediction tasks. The face is one of the easiest ways to distinguish the identity people. Face recognition is a type of personal identification system that employs a person’s personal traits to determine their identity. Human face recognition scheme generally consists of four steps, namely face detection, alignment, representation, and verification. In this paper, we propose to extract information from human face for several tasks based on recent advanced deep learning framework. The proposed approach outperforms the results in the state-of-the-art.

Corresponding Author:

Kittikhun Meethongjan
Department of Computer Science, Faculty of Science and Technology, Suan Sunandha Rajabhat University
1 U Thong Nok Rd, Dusit, Dusit District, Bangkok 10300, Thailand
Email: kittikhun.me@ssru.ac.th

1. INTRODUCTION

Face analysis is a major topic in machine vision, and it has been applied in various applications such as security surveillance, biometric recognition, tele-medicine, human behavior, kinship verification. Facial human can be used to extract different information: age or gender prediction, identification, and matching. Detecting human faces from a video is a challenging issue. Many advanced face detection and alignment approaches have been proposed in the past decades [1]–[3]. Early approaches of face analysis and recognition are based on the extraction of hand-crafted features [4]–[6]. For example, Vinay *et al.* [7] presents a double filter based on the extraction of GIST features for face recognition task. More recently, deep learning method prove its efficiency in computer vision with various applications [8]–[10]. Deng *et al.* [11] proposed a method for face detection based on the self-supervised learning combined with the extra-supervised method by using pixel-wise determination. Additive angular margin loss (ArcFace) [12] is a model proposed in 2019 for face recognition. This model also uses margin-based loss which outperforms other loss functions such as triplet loss from FaceNet. Multitask cascade convolutional neural network (MTCNN) [13] is a face detection and alignment method which aims at boosting both detection and alignment’s performance by exploiting the inherent correlation between the two processes. Li *et al.* [14] apply CNN cascade for improving the face detection stage. The CelebA [15] and WIDER FACE [16], [17] dataset is used for building training and evaluating this approach. Wang *et al.* [18] apply region-based fully convolution networks based on region-based fully convolutional networks (R-FCN) [19] for face detection. This architecture is applied for extracting features, and then fed into RPN to generate a batch of the region of interests (ROIs) according to the anchors. To aggregate the class scores and bounding box predictions, two global average pooling methods are applied to both class score maps and bounding box prediction maps in the final step. R-FCN is built upon ResNet-101 and consists of a region proposal network (RPN) and a R-FCN module. Deep hypersphere embedding for face recognition [20], [21] is a face recognition method proposed by Liu *et al.* [20]. The authors of SphereFace aim at improving the performance of face recognition model by implementing Angular softmax loss. In this paper, we propose to apply several recent advance deep learning frameworks for real-time face matching and gender prediction on videos. The rest of this paper is organized as: section 2 presents related works with face detection and alignment by RetinaFace and ArcFace. Section 3 introduces experimental setup and results. Finally, section 4 presents the conclusion and discuss the future works.

2. RELATED BACKGROUND

This section reviews face detection and alignment method, and generated face embeddings using ArcFace. After the face image is detected, the facial area is cropped and generated face embedding. These techniques are explained as follows:

2.1. Face detection and alignment

RetinaFace is achieved state-of-the-art performance by performing three different face localization tasks together, that are face detection, 2D face alignment and 3D face reconstruction based on a single shot framework. This model is robust as it achieved mean average precision (mAP) of 88.5 on WIDER FACE dataset. Figure 1 shows the detection and alignment procedure using RetinaFace. An image can be fed into this model to detect faces, the model then returns the facial area coordinates and facial landmarks (eyes, nose, and mouth). Consequently, the face can be extracted and aligned using these coordinates.

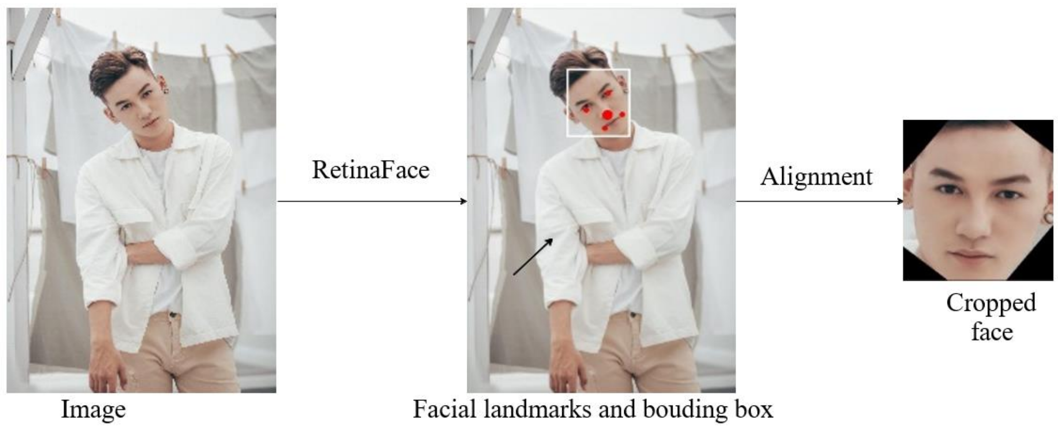


Figure 1. Face detection and alignment process by RetinaFace

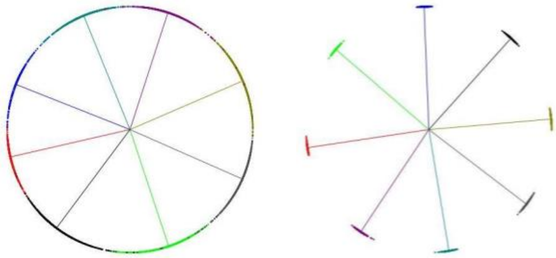
2.2. Generated face embeddings using ArcFace

Wang *et al.* [18] apply ResNet and improve this model by using smaller anchors and modify the position sensitive ROI pooling to a smaller size for suiting the detection of small faces. Next, they change the normal average pooling to position-sensitive average pooling for the last feature voting in R-FCN, which leads to improved embedding. Finally, multi-scale training strategy and online hard example mining (OHEM) strategy are adopted for training. Schroff *et al.* [22] introduced FaceNet which studied featured from facial images via a compact Euclidean space for enhancing the recognition and verification task. Zeiler and Fergus [23] investigated the performance of face recognition based on ImageNet dataset and large CNN models by a novel visualization approach. Moreover, ArcFace [12] is a model proposed by Deng *et al.* [11] for face recognition. For comparison, 8 different identities with enough sample (around 1,500 images/class) to train 2-D feature embedding networks with the softmax and ArcFace loss, respectively. Figure 2 shows examples of the softmax and ArcFace loss, the softmax loss generates notable ambiguity in decision boundaries but gives roughly separable feature embedding in Figure 2(a), whereas the suggested ArcFace loss may clearly enforce a larger separation between the neighboring classes in Figure 2(b).

Like SphereFace, ArcFace also uses margin-based loss which outperforms other loss functions such as triplet loss from FaceNet. ArcFace loss is based on softmax loss with modifications that give better discriminative power. These are formula of softmax loss and ArcFace loss respectively:

$$L_{\text{Softmax}} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{W_i^T x_i}}{\sum_{j=1}^{n} e^{W_i^T x_i}} \tag{1}$$

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{u}\left(\cos\left(\delta_{y_i}+m\right)\right)}{e^{u}\left(\cos\left(\delta_{y_i}+m\right)\right)+\sum_{j=1,j\neq y_i}^{n} e^{u}\cos\delta_j} \tag{2}$$



(a) (b)

Figure 2. Examples of 2D feature embedding networks (a) the softmax loss and (b) ArcFace loss

3. EXPERIMENTAL RESULTS

In this section, the experiment and the results are explained. Dataset are prepared for training and testing stage. Afterthat, the experimental results are shown and discussed.

3.1. Data preparation

Many datasets have been introduced [21], [24], [25] in the literature. In this paper, the VNCleb dataset is considered for evaluating the proposed approach. It consists of two parts: (1) the training subset has 21,626 face images of 100 celebrities. Each class in this subset has around 200 images. Figure 3 illustrates several images selected from this part, (2) the testing set contains 8,970 images of the above 100 celebrities and is cropped from 300 videos of these celebrities and was downloaded from YouTube with various resolutions. Each class in test set has around 90 images. Figure 4 shows the cropped images from the testing subset.

Figure 3. Selected images from the training set (1)

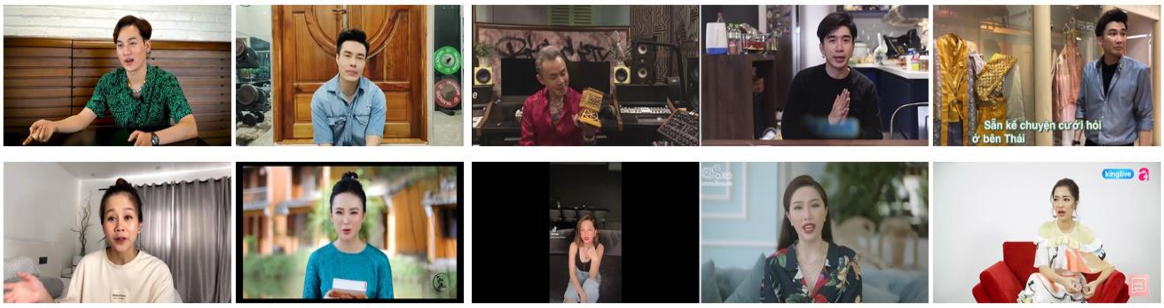


Figure 4. Selected images from the testing set (2). These images are cropped from the videos

Face images from testing subset follow the exact same pre-processing step which it is also represented by a 512-dimension array by using ArcFace. Several distance metrics, Euclidean, Manhattan and Cosine, are considered for comparing the two images. The process is illustrated in Figure 5. Figure 6 presents the scheme for gender and age prediction by using visual geometry group face (VGGFace) models. Face images are converted to 224×224 resolution. Images are then normalized by dividing each pixel by 255.

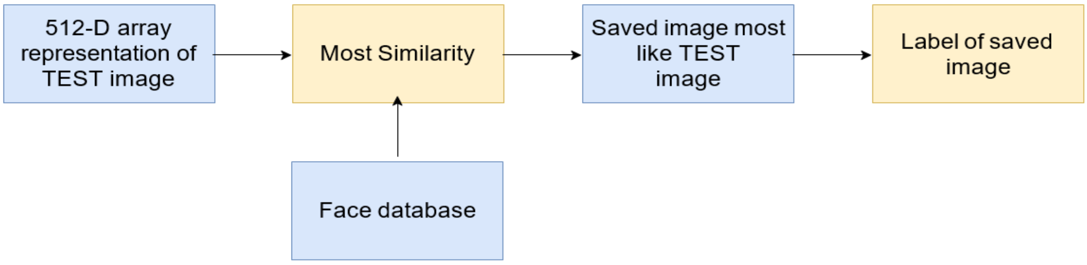


Figure 5. Perform face identification using distance metrics

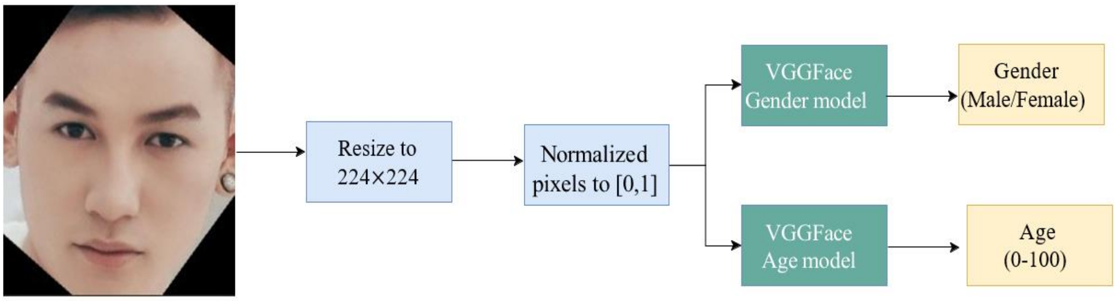


Figure 6. Gender and age prediction procedure using VGGFace

3.2. Results

This section presents the results achieved in experiments with different ArcFace models. We have tested 5 ArcFace models with a variety of backbones (ResNet18, ResNet34, ResNet50, ResNet100) pre-trained on different datasets (CASIA, Glint360k). The accuracy, precision, recall and F1 score metrics are employed to evaluate the performance. The performance of the proposed approach on the testing subset is summarized in Table 1. It can be learnt that training ArcFace on larger dataset like Glint360k (360 thousand class with a total of 17 million images) gives better result compare with the same model trained on CASIA (10 thousand class with a total of 0.5 million images). Furthermore, the accuracy is also affected by its architecture, as deeper ResNet architecture tends to outperform the shallower ones. ArcFace model with ResNet-100 backbone trained on Glint360k dataset outperforms others, achieving 94% accuracy, 0.93 on precision, 0.94 on recall and 0.93 on F1-score on testing subset.

Table 1. Performance comparison on different ArcFace models

Model	Parameters	Precision	Recall	F1	Accuracy
ArcFace+ResNet34+CASIA	34 million	0.82	0.82	0.80	0.82
ArcFace+ResNet34+Glint360k	34 million	0.86	0.87	0.85	0.86
ArcFace+ResNet100+Glint360k	65 million	**0.93**	**0.94**	**0.93**	**0.94**

ArcFace give a good performance on identifying the 100 celebrities of the testing subset with minimal false positive and false negative predictions. Confusion matrix of ResNet-100 ArcFace on the testing subset is illustrated in Figure 7. Investigating some images that is failed to predict, we observe that most incorrect predictions are given from faces that are turning right and left.

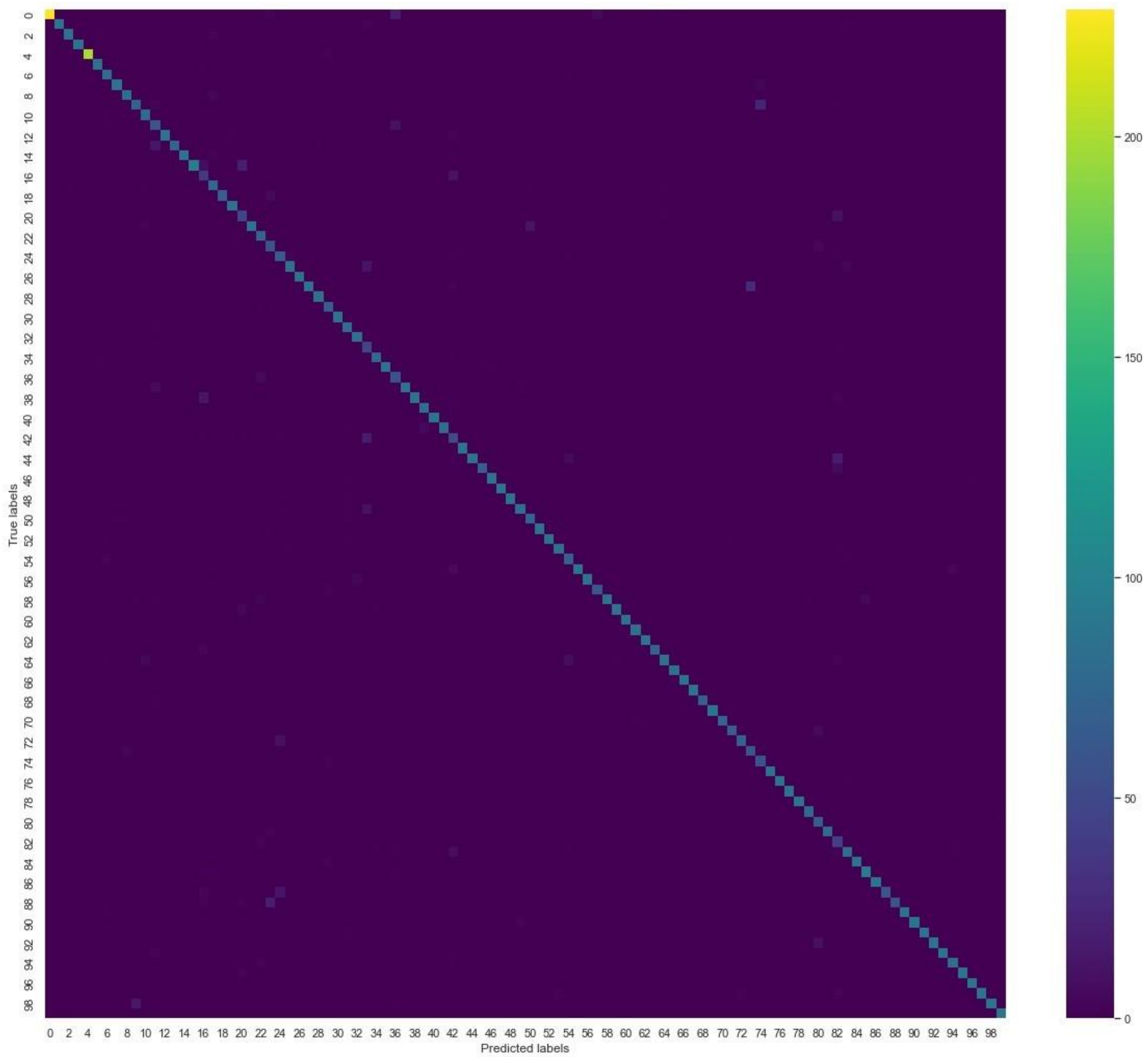


Figure 7. Confusion matrix of ResNet-100 ArcFace on the testing subset

Figure 8 shows several examples where ArcFace model fails to identify, most these are faces that are turned left/right to the point where we can only see half of the face. Therefore, model's performance improves when more face images of different position is added to the train set. Moreover, the gender and age estimation are predicted by using VGGFace model. The results are presented in Table 2. The VGGFace gender achives 94% of accuracy of gender prediction on the testing subset. Several failed cases are then selected to illustrate in Figure 9. We observe that these images only appear some part of faces so the model cannot detect face.

Figure 8. Several failed cases of the identification task. The model in used (from the top to bottom): ArcFace+R34+CASIA, ArcFace+R34+Glint360k, ArcFace+R100+Glint360k

Table 2. Performance of pre-trained VGGFace model on test set

Model	Parameters	Precision	Recall	F1	Accuracy
VGGFace Gender	134 million	0.94	0.94	0.94	0.94

Figure 9. Some failed cases of gender prediction task. Male and female on the first and second row, respectively

4. CONCLUSION

In this paper, we have applied RetinaFace for face detection and ArcFace for face identification. The ArcFace model with ResNet-100 backbone outperform other models as it has more layers, and it was trained on a very large dataset. While this model performs decently on the testing subset, there is still limitation as it is not performing well on side faces due to the lack of this kind of poses in the train dataset. We also applied VGGFace models for gender and age classification which has decent accuracy on the testing subset. The future of this work is now continuing to improve and compress model for better performance and representation.

ACKNOWLEDGEMENTS

This work was supported by Suan Sunandha Rajabhat University.

REFERENCES

[1] K. Meethongjan, M. Dzuikifli, P. K. REE, and M. Y. Nam, "Fusion affine moment invariants and wavelet packet features selection for face verification," *Journal of theoretical and applied information technology*, vol. 64, no. 3, 2014.
[2] A. Sobiecki, J. van Dijk, H. Folkertsma, and A. Telea, "Does face restoration improve face verification?," *Multimedia Tools and Applications*, vol. 80, no. 21–23, pp. 32863–32883, Sep. 2021, doi: 10.1007/s11042-021-11167-6.
[3] U. Saeed, K. Masood, and H. Dawood, "Illumination normalization techniques for makeup-invariant face recognition," *Computers and Electrical Engineering*, vol. 89, Jan. 2021, doi: 10.1016/j.compeleceng.2020.106921.

[4] J. Chen *et al.*, "Robust local features for remote face recognition," *Image and Vision Computing*, vol. 64, pp. 34–46, Aug. 2017, doi: 10.1016/j.imavis.2017.05.006.
[5] A. Essa and V. Asari, "Multi-texture local ternary pattern for face recognition," in *SPIE Proceedings*, May 2017, doi: 10.1117/12.2263735.
[6] W. Huang and H. Yin, "Robust face recognition with structural binary gradient patterns," *Pattern Recognition*, vol. 68, pp. 126–140, Aug. 2017, doi: 10.1016/j.patcog.2017.03.010.
[7] A. Vinay, B. Gagana, V. S. Shekhar, Anil B, K. N. B. Murthy, and S. Natarajan, "A double filtered GIST descriptor for face recognition," *Procedia Computer Science*, vol. 79, pp. 533–542, 2016, doi: 10.1016/j.procs.2016.03.068.
[8] S. Sethi, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread," *Journal of Biomedical Informatics*, vol. 120, Aug. 2021, doi: 10.1016/j.jbi.2021.103848.
[9] M. Wang and W. Deng, "Deep face recognition with clustering based domain adaptation," *Neurocomputing*, vol. 393, pp. 1–14, Jun. 2020, doi: 10.1016/j.neucom.2020.02.005.
[10] G. Chinnappa and M. K. Rajagopal, "RETRACTED ARTICLE: Residual attention network for deep face recognition using micro-expression image analysis," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. S1, pp. 117–117, Dec. 2022, doi: 10.1007/s12652-021-03003-4.
[11] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, May 2019.
[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 4685–4694, doi: 10.1109/CVPR.2019.00482.
[13] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, Jul. 2017, pp. 424–427, doi: 10.1109/ICISCE.2017.95.
[14] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 5325–5334, doi: 10.1109/CVPR.2015.7299170.
[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," Multimedia Laboratory, The Chinese University of Hong Kong. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed Aug 15, 2018).
[16] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 5525–5533, doi: 10.1109/CVPR.2016.596.
[17] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," *UMass Amherst technical report*, 2010.
[18] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," *arXiv preprint arXiv:1709.05256*, Sep. 2017.
[19] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.
[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6738–6746, doi: 10.1109/CVPR.2017.713.
[21] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv: 1411.7923*, Nov. 2014.
[22] 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.
[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision ECCV 2014*, Springer International Publishing, 2014, pp. 818–833.
[24] "YouTube-8M segments dataset," Research Google. https://research.google.com/youtube8m/ (accessed May 15, 2021).
[25] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

BIOGRAPHIES OF AUTHORS

Thongchai Surinwarangkoon received the B.Sc. degree in Mathematics from Chiang Mai University, Thailand, in 1995. He received the M.Sc. in management of information technology from Walailak University, Thailand in 2005 and Ph.D. in information technology from King Mongkut's University of Technology North Bangkok, Thailand in 2014. Currently, he is an assistant professor at Department of Business Computer, College of Innovation and Management, and the Deputy Dean for Administration of Graduate School, Suan Sunandha Rajabhat University, Bangkok, Thailand. His research interests include image processing, machine learning, artificial intelligence, business intelligence, and mobile application development for business. He can be contacted at email: thongchai.su@ssru.ac.th.

Vinh Truong Hoang received his master's degree from the University of Montpellier in 2009 and his Ph.D. in computer science from the University of the Littoral Opal Coast, France. He is currently an assistant professor, Head of Data Science Department, and the Vice Dean of Faculty of Information Technology, Ho Chi Minh City Open University in Vietnam. His research interests include deep learning, feature selection, texture classification, local binary patterns, and face recognition. He can be contacted at email: vinh.th@ou.edu.vn.

Ali Vafaei-Zadeh received bachelor's degree in industrial management from Allameh University, Iran. He received MBA in production and materials management from University of Pune, India, and the Ph.D. in operation management from Universiti Sains Malaysia, Malaysia. Currently, he is a senior lecturer at the Graduate School of Business (GSB), Universiti Sains Malaysia (USM). He is actively involved in research projects regarding technology management and innovation. His research interests include technology and innovation management. He can be contacted at email: vafaei@usm.my.

Hayder Ibrahim Hendi is an assistant Professor at University of Thi Qar, the Computer and Math College. Hendi was born in naissrayah in 1976. He received his master of computer sciences from informatics high studies in states, Iraq 2006. In 2017, he received his Ph.D. degree in information and computer sciences from ULCO unversite (calais, France). His many researchs that interests lie in the field of security, ontology, and semantics web, logistics system and optimization. Recently, his research interests included IoT and optimization. He can be contacted at email: dr-hayder@utq.edu.iq, hayder.ibrahim8@gmail.com.

Kittikhun Meethongjan received his B.Sc. degree in computer science from Suan Sunandha Rajabhat University, Thailand, in 1990, a master's degree in computer and information technology from King Mongkut's University of Technology Thonburi (KMUTT), Thailand, in 2000, and the Ph.D. in computer graphic from the University of Technology Malaysia in 2013. Currently, he is an assistant professor at the Computer Science Program, Head of Apply Science Department, Faculty of Science and Technology, Suan Sunandha Rajabhat University (SSRU), Thailand. His research interests include computer graphic, image processing, artificial intelligent, biometrics, pattern recognition, soft computing techniques and application. He can be contacted at email: kittikhun.me@ssru.ac.th.