ICIC Express Letters
Volume **15**, Number **1**, January **2021**

 ISSN 1881-803X
pp. **57–65**

FACE RECOGNITION FOR WORK ATTENDANCE USING MULTITASK CONVOLUTIONAL NEURAL NETWORK (MTCNN) AND PRE-TRAINED FACENET

Fadhillah Moulita Andiani and Benfano Soewito

Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
JL. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
fadhillah.moulita@binus.ac.id; bsoewito@binus.edu

Received February 2020; accepted May 2020

Abstract. *Attendance has become one of the most important needs in any area, especially in workplaces. Traditional attendance by using handwritten signatures has proven to be very inefficient due to its challenging instances of time-consuming, falsifications, impersonations and miscalculations. In addition to the traditional way, among other approaches used for recording attendance, biometric technique is the one attendance method. Using biometric attendance, such as fingerprint, voice, and face will be simplify the process of attendance, and employee does not need to sign on attendance sheets every day. This technique is more efficient and saves times than traditional way. This paper proposed a face recognition as biometric attendance system using deep learning method. The entire process of developing a face recognition model will be described in detail. This face recognition model is composed of several essential steps using the most advanced technique: Multitask Convolutional Neural Network (MTCNN) for face detection and FaceNet for embeddings – verification process to recognize face. This goal must be met as a result of the attendance system, and every condition must be concerned to return the valid and accurate result. This model will gather the face image of the employee and will return 128-d embedded vectors. Thus, result of the attendance process will be stored to excel sheet and the system will automatically mark the attendance once employee is detected by camera. After implementing the model, overall accuracy was 95% on a small dataset of the original face images of employees in the real-time environment.*
Keywords: Face recognition, Deep learning, Attendance, Multitask convolutional neural network

1. **Introduction.** Recording hours of work for workers is one of a company's important components. This process will also calculate the performance and the remaining days off in addition to storing the attendance. When handled manually, the attendance reporting process will take a lot of times. As a result of a rapid growth in information technology, automatic attendance has become an option for these types of business processes, particularly in Internet of Things (IoT). There are plenty of system that can be implemented to achieve higher accuracy of attendance record, and one of the methods is biometric authentication. Biometric authentication also offers a number of ways to verify the identity of the person, such as using fingerprints, iris, voice, and face [1-4]. Recently, face recognition became the most popular area of research and study in computer vision. This method offers many opportunities and challenges that need to be explored. An approach using *Multitask Convolutional Neural Network* (MTCNN) joints the process of detection and alignment to improve accuracy with a fast run time when face is recognized [5]. *Multitask Convolutional Neural Network* (MTCNN) approach uses three-stage multi-task deep convolutional networks, and the final network will produce final bounding box and facial landmarks position. The goal of this study is to explore new advanced techniques using

DOI: 10.24507/icicel.15.01.57

in-depth learning to overcome limitations such as low precision when detecting multiple faces, and intense light conditions. *Multitask convolutional neural network* approach will be used to detect employee's faces. To resolve limitation of detecting multiple faces, the result of detection will return back multiple sets of coordinates and finally author's system will iterate through every individual face and drew out the bounding box frame and dot the 5 facial landmarks. For face embeddings – face verification using FaceNet, this research will train the FaceNet model with small employee's dataset and will load into the attendance system.

The contribution of this research is to provide a solution for face recognition tasks and create the most efficient tool for manipulating the attendance system, using the proposed approach. The rest of the paper is organized as follows: Section 2 presents the related work, Section 3 presents the preparation and methodology of results, Section 4 presents the outcome and discussion, and the final section holds the conclusion.

2. **Related Work.** There are now many types of computerized monitoring and attendance systems using face recognition applied in organizations as a result of active research in information technology. Modular Fisher Linear Discriminating Analysis (M-FLDA) implemented by Lang and Hong [6] checked the entrance guard and work attendance based on face recognition. This approach divides images into blocks of images and selects subimages according to the new lower-dimensional pattern, and the analysis method is used based on image segmentation [7-9]. The advantages of this entire process are to efficiently remove the local distinguishing features of the original image, reduce memory capacity, and also return 90.83% accuracy for 360 number of test samples. However, the recognition rate is reduced when this research improves the recognition speed. Hongo detected multiple faces and hand recognition using multiple camera [10]. Those cameras are made up of 2 image trackers and 2 stereo cameras. The stereo camera was used to detect faces and determine increasing distance from the stereo camera to be processed by a quadruple image transfer, while the tracking camera was used to zoom and tilt angles of faces. The accuracy was 96%, but it will decrease when people overlapping. Arsenovic et al. [11] developed attendance using face recognition and deep learning, and CNN cascade is used as a face detector [12-14]. This model also used FaceNet to classify faces as facial embedding and Support Vector Machine (SVM) classifier to classify faces. The drawback was that this model is not applied in many faces and when intense lighting conditions the accuracy decreased but still provided high accuracy around 95.02%. Zhang and Zhang [15] also tested under controlled condition for recognized face. Zhang and Zhang presented recognized face when the condition was under controlled such as pose, intense lightning and face expression, and it is combined multiple resolution for under controlled condition when detecting face using CNN cascade [16]. This research involves face detection boundary box calibration with additional cost of computation.

Zhang et al. [5] detected multiple faces, face expression, and overlapping condition using Convolutional Neural Network (CNN) cascade with multitask learning. This approach is integrated face detection and face synchronization as a result of which the state-of-the-art methods are consistently surpassed by several daunting benchmarks [17-20]. Using FDDB [21] and WIDER FACE dataset benchmarks for face detection and AFLW dataset benchmarks for face alignment also keeps real-time performance but needs to exploit the inherent correlation between face detection and other face analysis tasks to further improve the performance. Taigman et al. [22] trained FaceNet embeddings using LFW dataset. It presented embeddings using 128-d and using one shot model directly to map from face to Euclidean space [23,24]. Accuracy of this method was 99.63% but it improves error rate when processing extra face alignment.

By the past research, multiple faces detection in attendance system has been studied but still has limitation like the reduced recognition rate, decreased accuracy when people

overlapping and intense lightning conditions. The limitation motivated authors to use an alternated version of this approach as part of model for the deep learning using multitask based face recognition attendance system implemented in workplaces.

3. **Material and Methods.** The process of developing the multitask convolutional neural network-based attendance system will be explained in detail in this section. Developing procedure is divided into several important stages, including obtaining the training dataset, preparing images and training MTCNN and last is integration into attendance system in order to test the proposed method.

3.1. **Employee data preparation.** The implemented system in this paper was tested in a company and ten employees volunteered in this research. All employees stored their photos and information to employee database, it took several different positions while being photographed. Total of this small dataset is about 1030. Result will return 128-d embedded vector $[-0{,}112, 0{,}9312, \ldots, n]$ and this vector will be used as verification process.

3.2. **Dataset preparation and training.** Since MTCNN composed of 3 separate neural networks that could not be trained together, authors decided to start by training P-Net as the first network. This model used WIDER FACE dataset to train bounding box coordinates and CelebA dataset to train face landmarks. For simplicity, authors started by training only the bounding box coordinates. The WIDER FACE [23] dataset includes 32,203 images with 393,793 faces of people in different situations.

3.3. **Developing face recognition model.** This model contains several important steps: face detection and face landmarks using multitask convolutional neural network, generating face embeddings and face classification, and picture below shows the face recognition model.

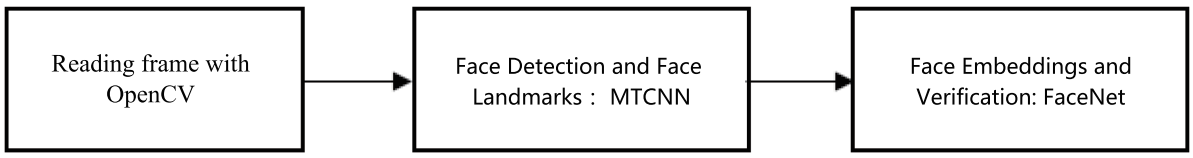

FIGURE 1. Face recognition model

The first step of the face recognition process is face detection. Face detection presents from well-studied field in computer vision domain. Face alignment also attracts extensive interests. Regression based methods [24] and template fitting approaches [25] are two popular categories. As a result of decades of research, there are numerous machine learning algorithms applicable for this task. Recently, Convolutional Neural Network (CNN) cascade with multitask learning has been developed by Zhang et al. [5]. This method is used to face detection and face landmarks, this model works in three steps, and Figure 2 shows the architecture of MTCNN.

Multitask cascade convolutional neural network has 3 steps to detect face and generate face landmark.

Step 1: First step is fully convolutional network, called P-Network (P-Net) and used to predict face positions and bounding box. This P-Net created image pyramid in order to detect faces of all different sizes. In other words, this network wants to create different copies of the same image in different sizes to search for different sized faces within the image. Each copy of scaled images, will scan faces using 12×12 stage 1 kernel in every part of the image. It starts in the top left corner, a section of the image from $(0, 0)$ to $(12, 12)$. This portion of the image is passed to P-Net, which returns the coordinates of a bounding box if it notices a face. Then, it would repeat that process with sections

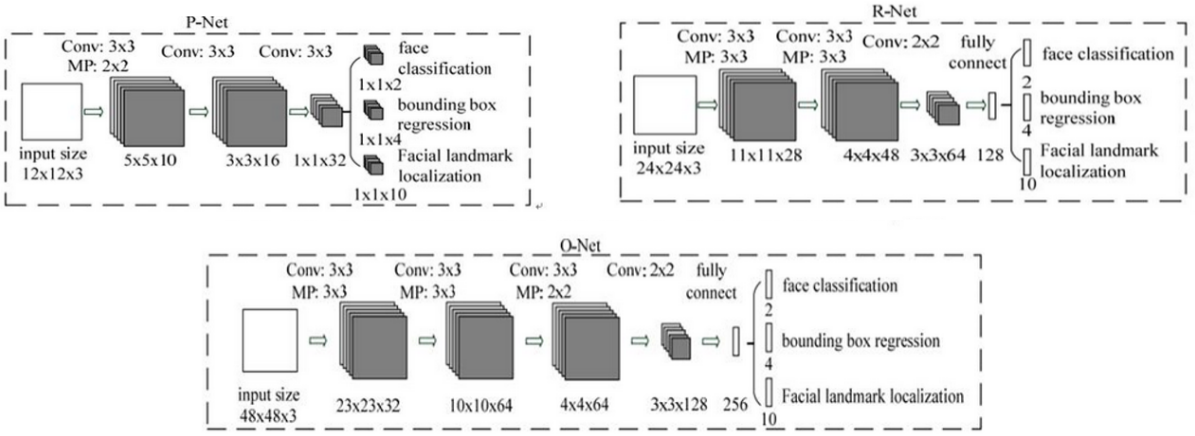

FIGURE 2. MTCNN architecture

$(0+2a, 0+2b)$ to $(12+2a, 12+2b)$, shifting the 12×12 kernel 2 pixels right or down at a time. The shift of 2 pixels is known as the **stride**, or how many pixels the kernel moves by every time. This stride of 2 helps to reduce computation complexity without significantly sacrificing accuracy. After passing the multiple scaled copies of the image and gathering its output, authors need to parse the P-Net output to get a list of confidence levels for each bounding box and delete the box with lower confidence.

Step 2: Sometimes an image may contain only a part of face peeking in from the side of the frame. The second step is to solve that case using R-Net. For every bounding box, this network creates an array of the same size and copies the pixel values (the image in the bounding box) to the new array. If the bounding box is out of bounds, R-Net copies the portion of the image in the bounding box to the new array and fills in everything else with a 0 (padding). In image below, the new array for McCartney's face would have pixel values in the left side of the box, and several columns of 0s near the right edge.

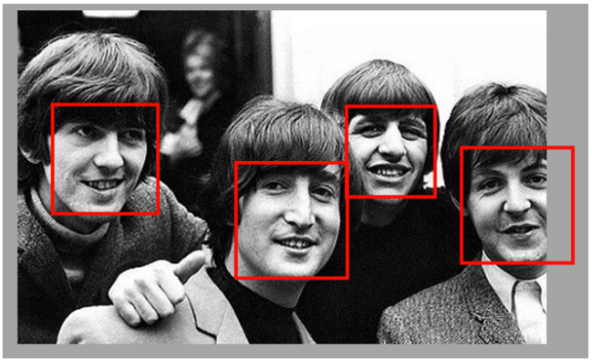

FIGURE 3. Images with bounding boxes

Image above shows the beetles has bounding boxes in that picture. After padding the bounding box arrays, it needs to resize to 24×24 pixels, and normalize them to values between -1 and 1. Currently the pixel values are between 0 to 255 (RGB values). By subtracting each pixel value by half of 255 (127.5) and dividing it by 127.5, this network can keep their value between -1 and 1, after that we can feed them into R-Net and gather its output. R-Net output is similar to that of P-Net: It includes the new accurate bounding boxes, as well as the confidence level of each of these bounding boxes.

Step 3: Before passing in the bounding boxes from R-Net, we have to first pad any boxes that are out-of-bounds. Then, after we resize the boxes to 48×48 pixels, we can pass in the bounding boxes into O-Net. The outputs of O-Net are slightly different from that of P-Net and R-Net. O-Net provides 3 outputs: the coordinates of the bounding box

(out [0]), the coordinates of the 5 facial landmarks (out [1]), and the confidence level of each box (out [2]). Once again, we get rid of the boxes with lower confidence levels and standardize both the bounding box coordinates and the facial landmark [26].

The second step is embeddings and verification process using FaceNet. FaceNet is a deep convolutional neural network which is developed by Google researchers [27] and introduced around 2015 to effectively solve the hurdles in face detection and face verification. Algorithm of FaceNet transforms the face image into 128-dimensional Euclidean space similar to word embedding, also this model trained for triplet loss to capture the similarities and differences on the image dataset provided. Using FaceNet embeddings as future vectors, functionalities such as face recognition, and verification could be implemented after creating the vector space [28]. Thus, the distances for the similar images would be much closer than the random non similar images. Picture below shows the architecture of FaceNet model.

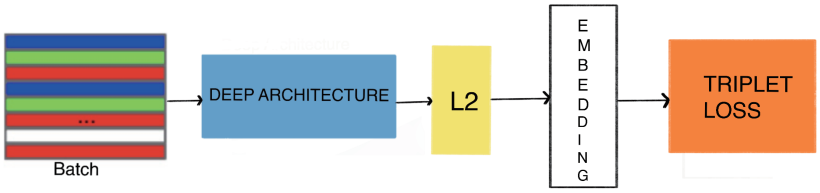

FIGURE 4. FaceNet architecture

FaceNet network architecture consists of a batch input layer and deep convolutional neural network, network followed by L2 normalization, that provides the face embeddings. This process is in turn followed by the triplet loss [29]. Triplet loss is proven to be very effective for face verification/recognition and also related domain such as person re-identification [30]. By enforcing a margin between the pairs of faces of the same identity and the ones of the different identities, the triplet loss tries to keep the faces of the same identity closer than the faces from the different identities in the embedding space. This allows the faces for one identity to live on a manifold while still enforcing the distance and thus discriminating to other identities. Triplet loss can be calculated using formula below:

$$\|f\left(x_i^a\right) - f\left(x_i^p\right)\|_2^2 + \alpha < \|f\left(x_i^a\right) - f\left(x_i^n\right)\|_2^2 \tag{1}$$

$f(x_i)$ represents the embedding of an image, x_i represents an image, and α represents the margin between positive and negative pairs. The superscripts a, p and n correspond to anchor, positive and negative images respectively. α is defined here as the margin between positive and negative pairs. It is essentially a threshold value which determines the difference between our image pairs. If alpha is set to 0.5, then the difference between anchor positive and anchor negative image pairs is to be at least 0.5. Thus, triplet loss is one of the best ways to learn 128-dimensional embedding for each individual face better than other way.

4. **Result Analysis.** Train MTCNN model to automatically create face landmark and use the pretrained FaceNet model to create the embedding. Then real-time face detection and face recognition are performed, which include the following sub processes: reading frames with OpenCV, face detection and landmark using MTCNN, face embedding and verification using FaceNet algorithm, and storing the recognized face for analysis.

Figure 5 shows the example result of face detection in multiple faces, and the labels of bounding box show the name of employee and accuracy of face recognition.

The identified face is exported into an excel sheet for registering the presence of student, and the excel sheet that displays the demo output of attendance is shown in Figure 6.

FIGURE 5. Implementation result of face detection

Name	Roll Number	11_02_20
Della	11223344	Present
Vega	827283	Present
Grace	82937391	Present
Gaga	12839374	Present

FIGURE 6. Excel sheet with sample output

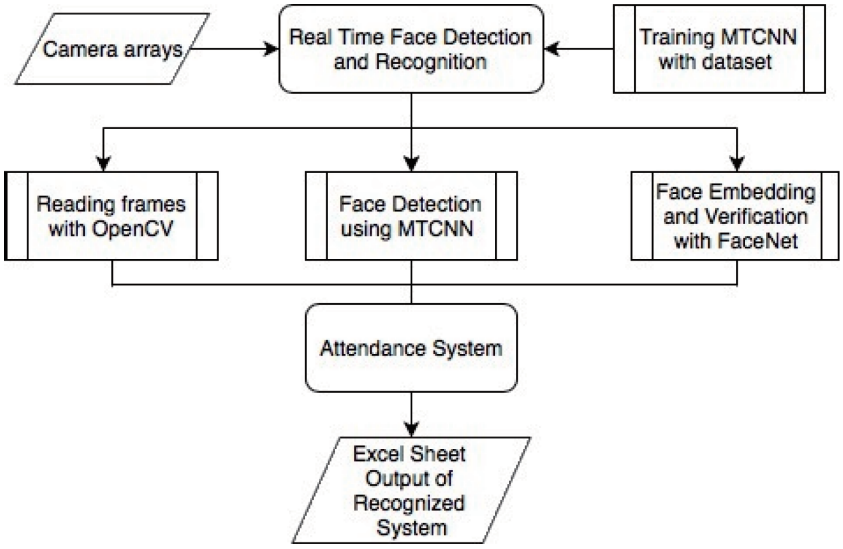

FIGURE 7. Block diagram of implementation

Figure 7 shows the block diagram of implementation face recognition model using MTCNN and FaceNet, which camera arrays as input in this system and excel sheet of attendance as output. To measure the performance of the model, authors set threshold to 70% where this value is got from error equal rate of pre-trained FaceNet model. This research also used confusion matrix to assess the performance of classification [31].

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{total testing image}} \tag{2}$$

$$\text{Precision} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}} \tag{4}$$

$$\text{F-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The precision, recall and f-measure for face detection evaluation can be calculated as follows, where TP (True Positive) is a detection bounding box presented in System Under Test (SUT) and Ground Truth (GT), FP (False Positive) presents in SUT but not in GT and FN (False Negative) presents in GT but not in SUT. This model tested about 40 times with random total multiple person (above 6 person) and random different lightning condition. Table 1 provides the result of confusion matrix of the model and can be analyzed that the system is having an accuracy of 95% for detection in multiple faces. 0 shows true values and 1 shows false values. Precision has values above 90% in both conditions. Recall has a value above 90%. This signifies the true positive rate of the system. F1-score of the known classes is also above 90% and has maximum value above 94%. This shows that there is good balance between precision and recall.

Table 1. Confusion matrix for multiple person detection

	Precision	Recall	F1-score
0	0.94	0.94	0.94
1	0.95	0.95	0.95
micro avg	0.95	0.95	0.95
macro avg	0.95	0.95	0.95
weighted avg	0.95	0.95	0.95

Table 2 provides the result of confusion matrix of the model and can be analyzed that the system is having an accuracy of 84% for multiple faces detection in different lightning condition. This value has significant decrement for this case, total of difference above 10%.

Table 2. Confusion matrix for under controlled condition

	Precision	Recall	F1-score
0	0.84	0.89	0.86
1	0.90	0.86	0.88
micro avg	0.88	0.88	0.88
macro avg	0.87	0.88	0.87
weighted avg	0.88	0.88	0.88

Both of testing are giving the high values, but when testing on multiple faces in under controlled condition return lower values in precision, recall, and f-score than normal detection because some of the results give unknown values when employee detected in extreme lightning condition. This proposed method returns 95% of accuracy when detecting multiple faces, and this result still provided the high accuracy. Arsenovic et al. [11] proposed method using CNN for detection, FaceNet for face embeddings, and support vector machine for face classification, has accuracy around 95% but that proposed method cannot be able to detect multiple faces in one frame.

5. **Conclusion.** Nowadays, various attendance and monitoring tools are used in practice industry. Regardless the fact that these solutions are mostly automatic, they are still prone to errors. In this paper, a new framework for real-time multiple face recognition using MTCNN and FaceNet as attendance system is proposed. This standalone system detects the person which was already given in the dataset to track and embedding being created was successfully detected with an accuracy of 95%. When testing on multiple faces in under controlled condition return false detection and lower values because using one shot learning, the future work could involve exploiting newly gathered images

in runtime for automatic retraining of the embedding FaceNet. One of the unexplored areas for this research is analysis of additional solutions for classifying face embedding vectors. Developing a specialized classifying solution for this task could potentially lead to achieving higher accuracy on a smaller dataset. This deep learning model does not depend on GPU in runtime, and it could be applicable in many other systems as a main or a side component that could run on a cheaper and low-capacity hardware, even as a general-purpose Internet of Things (IoT) device.

REFERENCES

[1] A. K. Jain, A. Ross and S. Pankanti, Biometrics: A tool for information security, *IEEE Trans. Information Forensics and Security*, vol.1, no.2, pp.125-144, 2006.

[2] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, Springer Verlag, New York, 2004.

[3] D. Maltoni, D. Maio, A. K. Jain and S. Prabhakar, *Handbook of Fingerprint Recognition*, Springer Verlag, New York, 2003.

[4] J. Daugman, The importance of being random: Statistical principles of iris recognition, *Pattern Recognit.*, vol.36, no.2, pp.279-291, 2003.

[5] 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.11, pp.1499-1503, 2016.

[6] L. Lang and Y. Hong, The study of entrance guard & check on work attendance system based on face recognition, *International Conference on Computer Science and Information Technology*, 2018.

[7] X. Tan, J. Liu and S. Chen, Sub-intrapersonal space analysis for face recognition, *Neurocomputing*, pp.1796-1801, 2006.

[8] J. R. Price and T. F. Gee, Face recognition using direct, weighted linear discriminant analysis and modular subspaces, *Pattern Recognition*, vol.38, pp.209-219, 2005.

[9] J. Yang, J.-Y. Yang and A. F. Frangi, Combined fisherfaces framework, *Image and Vision Computing*, vol.21, pp.1037-1044, 2003.

[10] H. Hongo, Focus on attention for face and hand gesture recognition using multiple cameras, *IEEE International Conference on Automatic Face and Gesture Recognition*, 2009.

[11] M. Arsenovic, S. Sladojevic, A. Anderla and D. Stefanovic, FaceTime – Deep learning based face recognition attendance system, *IEEE International Symposium on Intelligent System and Informatics*, 2017.

[12] H. Li et al., A convolutional neural network cascade for face detection, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[13] O. Russakovsky et al., ImageNet large scale visual recognition challenge, *International Journal of Computer Vision*, vol.115, pp.211-252, 2015.

[14] X. Zhu and D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] C. Zhang and Z. Zhang, Improving multiview face detection with multi-task deep convolutional neural networks, *IEEE Winter Conference on Applications of Computer Vision*, pp.1036-1041, 2014.

[16] X. Xiong and F. Torre, Supervised descent method and its applications to face alignment, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.532-539, 2013.

[17] Z. Zhang, P. Luo, C. C. Loy and X. Tang, Facial landmark detection by deep multi-task learning, *European Conference on Computer Vision*, pp.94-108, 2014.

[18] Z. Liu, P. Luo, X. Wang and X. Tang, Deep learning face attributes in the wild, *IEEE International Conference on Computer Vision*, pp.3730-3738, 2015.

[19] V. Jain and E. G. Learned-Miller, *FDDB: A Benchmark for Face Detection in Unconstrained Settings*, Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010.

[20] F. Schroff, D. Kalenichenko and J. Philbin, FaceNet: A unified embedding for face recognition and clustering, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.815-823, 2015.

[21] Y. Sun, X. Wang and X. Tang, Deeply learned face representations are sparse, selective, and robust, *CoRR*, abs/1412.1265, 2014.

[22] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, DeepFace: Closing the gap to human-level performance in face verification, *IEEE Conference on CVPR*, 2014.

[23] *WIDER FACE Database*, Mmlab.ie.cuhk.edu, 2019.

[24] X. Cao, Y. Wei, F. Wen and J. Sun, Face alignment by explicit shape regression, *International Journal of Computer Vision*, vol.107, no.2, pp.177-190, 2012.

[25] T. F. Cootes, G. J. Edwards and C. J. Taylor, Active appearance models, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.23, no.6, pp.681-685, 2011.

[26] Y. Sun, Y. Chen, X. Wang and X. Tang, Deep learning face representation by joint identification-verification, *Advances in Neural Information Processing Systems*, pp.1988-1996, 2014.

[27] Google FaceNet scores almost 100% recognition, *Biometric Technology Today*, vol.2015, no.4, pp.2-3, 2015.

[28] S. D. Shendre, An efficient way to trace human by implementing face recognition technique using TensorFlow and FaceNet API, *International Journal for Research in Applied Science and Engineering Technology*, vol.6, no.4, pp.605-608, 2018.

[29] Z. Ming, J. Chazalon, M. M. Luqman, M. Visani and J.-C. Burie, Simple triplet loss based on intra/inter-class metric learning for face verification, *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[30] D. Cheng, Y. Gong, S. Zhou, J. Wang and N. Zheng, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1335-1344, 2016.

[31] A. Baumann, A review and comparison of measures for automatic video surveillance systems, *EURASIP Journal on Image and Video Processing*, vol.2008, no.1, p.824726, 2008.