An Efficient Face Detection and Recognition

Thanh Tan Nguyen Thi, Khanh Nguyen Trong

***Abstract*—In this article, we propose a new method to effectively recognize faces from connected devices like real-time camera or webcam. The method contains two phases: Detecting and recognizing faces from the webcam frame. The face detection phase uses HOG features and SVM linear classifier. The second phase bases on FaceNet neural network model to automatically extract facial features and SVM classifiers. The experiments with UOF, FEI, JAFFE and LZW dataset is presented to show the efficiency of the proposed method. Experimental results show that the proposed method achieves high accuracy and stability on the test data sets collected from the actual environment.**

***Index Terms*— Face recognition, Real-time recognition, Frame based recognigition, Recognition deep neural network.**

I. Introduction

The biometric and also face recognition problem has been attracted many researches since last decade. Actually, those technologies are used not only for personal identification but also for many practical problems such as access control, network access control, in important areas such as terminals, airports, banks, time attendance automation, etc.

In Vietnam, the biometric technology has also been widely used in many domains, such as automatic time attendance systems based on fingerprint or face recognition, security monitoring systems, detection objects, intrusion detection, detection and alerts, abnormalities … However, most of used technologies are from closed and does not support real-time processing.

We are interesting in an overall solution for real-time monitoring systems based on face recognition. The proposed solution contains two phases: (i) direct face detection from video frames; and (ii) face recognition based on deep learning approach. The detection method is based HOG features and linear SVM classifier [11]. For the recognition method, we combine SVM classifier method with FaceNet deep neural network model [5] that automatically extracts facial features. The paper is organized as following: Section 2 present related works in face recognition; section 3 present our method; in the section 4, we show experiments of our method with common dataset in the field; section 5 is our conclusion and future work.

The remainder of this paper is organized as follows. In Section 2, relevant research works are reviewed. In Section 3, the designing and development of CIDE are discussed. Some discussions will be presented in section 4. Finally, the conclusion of this work is given in Section 5.

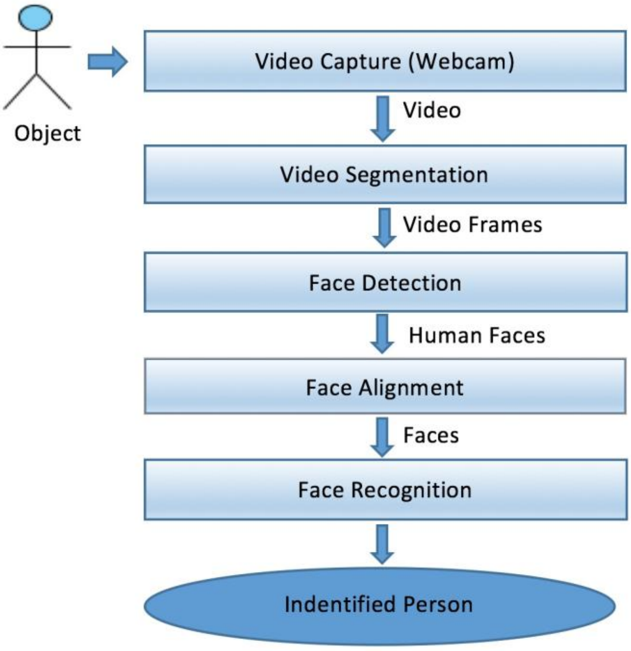


Figure 1. Real-time face recognition from webcam algorithm

II. Related works

Face recognition is the process of identifying an individual subject automatically in a photo/video based on content. Many approaches have been proposed to solve this problem [7], [9], [15], [8]. In general, the process usually consists of the following steps: (i) Image acquisition; (ii) pre-processing, enhancing image quality; (iii) Detecting, aligning, cropping faces; (iv) Face Identification (extraction and characterization) of the face.

In the literature, many approaches are feature-based and usually bases on distance, area, and angle to provide explicit definitions for facial expressions [15]. An explicit facial expression allows to have a visual space. However, in practice, explicit expressions are often inaccurate. To overcome this difficulty, many researches applied the statistical machine learning method which can learn to select facial features from a given set. For instance, Principal Component Analysis (PCA) method, in which each face is represented by a combination of eigenvectors, Eigen faces and fisher faces [10], [17], or Convolutional Neural Network [16].

Actually, the performance of face detection model has been significantly improved by the combination of deep learning model to automatically detect facial features and statistical classification technologies.

Revised Manuscript Received on February 2018.

Dr. Thanh Tan Nguyen Thi, Computer Science and Information Systems Department, Information Technology Faculty, Electric Power University, in HaNoi, VietNam (email: tanntt@epu.edu.vn)

Dr. Khanh Nguyen Trong, Software Engineering, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam, (84 4) 912314482., (e-mail: khanhnt@ptit.edu.vn).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

In [20], [21], and [22], the authors proposed a multi-stage model that bases on the combination outputs of a deep convolutional neural network D-CNN with PCA to reduce data dimension, and the SVM classifier.

Zhenyao *et al.* [22] constructed a deep neural network to align faces and then trained a CNN to classify and identify each face Y. Taigman *et al.* [21] proposed a model, named DeepFace that is a multi-stage integration. Firstly, the authors use a 3-D face model to standardize input images (collected with different angles). Secondly, they build a deep neural network (DNN) with 120 million parameters that is capable of learning from a huge database of over 4.4 million labeled faces. In the DeepFace DNN network, the last network layer is removed and the output of the previous network layer is used as a low-dimensional representation of face. The empirical results show that the model has an accuracy of 97.35% for the LFW dataset [6].

In general, as many other methods, DeepFace represented faces by low-dimensional representations that is generalized for new faces that the network has never trained. However, to do this, the method needs to train with a huge dataset that requires a high performance infrastructure.

In [5], Florian Schroff *et al.* proposed a neural network architecture, named FaceNet, with the triplet loss function that is defined directly on the representation. Figure 1 illustrates how FaceNet's training procedure learns to cluster face representations of the same person. The unit hypersphere is a high-dimensional sphere such that every point has distance 1 from the origin.

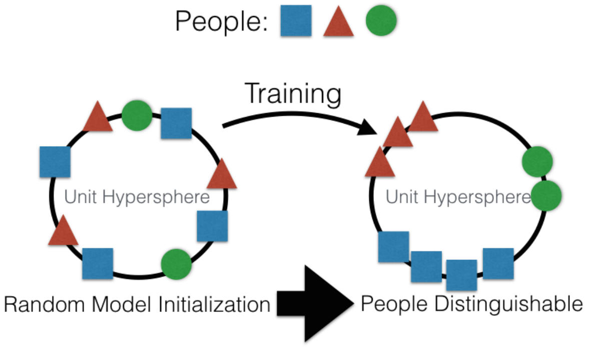


Figure 1. FaceNet training and triplet-loss function

Important improvements of FaceNet include: (i) Application of cost triplication; (ii) Selection of triplet during training; (iii) Allowing learning from huge data sets to find the appropriate network architecture.

III. METHODOLOGY

In fact, the identification of objects and also face recognition directly from the camera or webcam system or are challenging tasks owing to their variable appearance and the wide range of poses that they can adopt. One of the typical challenges is that the object's face is usually motive and changed in the posture, angle, or behavior. It requires the recognition algorithms to be generalize that is not affected much by those factors. In addition, direct recognition from the camera/webcam also requires real-time response. Therefore, we propose a real-time face recognition algorithm from such systems, which have 6 steps, as shown in Figure 2:

- Video Capture: By a webcam that connect to our application, we can get video about object.
- Video Segmentation: from the output of previous step, this step segments videos into separate frames. The segmentation is performed following time factor with threshold 24 frames per second.
- Face Detection: Each frame may not contain, partly contain, or entirely contain face. So that, in this step, the algorithm will detect and locate face on the image, if any.
- Face Alignment: The frame that is successfully detected faces will be applied some pre-processing to enhance image quality, such as reducing noise, de-frosting/blurring, standardization of size and resolution, face alignment to face-to-face look.
- Face Recognition: Faces after pre-processing will be used as inputs for a deep neural network model (DNN). The model will automatically learn and extracts features to identify (classification) the face.
- Identification: The final step of the algorithm will classify (identify) the faces. The face classification is to find the person whose face pattern is most similar to the face that needs identification. To accomplish this, the layered models need to be trained with a given sample set. Each face pattern is represented by a set of features obtained from the DNN detection model in the previous step.

A. *Face detection*

As mentioned earlier, the nature of face detection is finding and locating faces on any photo frame. In the proposed method, we use the HOG (Histograms of Oriented Gradients) and Support Vector Machines [11] for face detection.

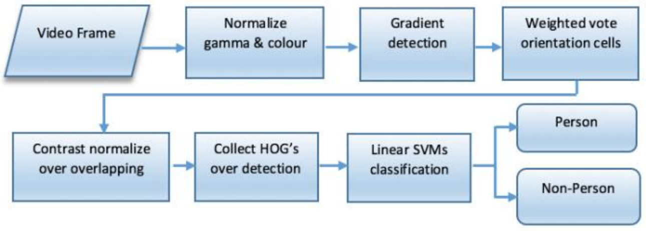

Video Frame
Normalize gamma & colour
Gradient detection
Weighted vote orientation cells
Contrast normalize over overlapping
Collect HOG's over detection
Linear SVMs classification
Person
Non-Person

Figure 2. Face detection algorithm

The main idea of the HOG feature is that object's shape and state can be featured by the gradient distribution and edge direction. The feature is developed based on the Scale-Invariant Feature Transform (SIFT) feature, which is based on the HOG feature. Because of the color variation in different regions, therefore each region has its own feature vector. So in order to get the features of the whole window, we have to combine many consecutive areas together. These steps are described in detail in Figure 3.

The input of the algorithm is frame obtained from the segmentation step. The obtained frame (RGB image) is converted into a gray scale image that is then applied histogram balance to reduce the light sensitivity. In the next step, the algorithm will calculate the color variation of all pixels of gray scale image in X [-1, 0, 1] and in Y $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$ direction.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

The result of this step is two gradient-x and gradient-y images that have the same size as the gray scale image. Then, the images are processed to calculate the angle and direction of the color variation.

The precise storing of the orientation of each pixel (x, y) is cost-consuming and not efficient, so we divide the angle space into bin. The smaller bin divides, the greater accuracy is. The empirical results [18] shows that with the bin size 200 gives, we obtain the best result of face detection. Therefore, with the space direction from 0^0 to 180^0, it will be divided into 9 bin as follows: [0^0 - 20^0], [21^0 - 40^0], [41^0 - 60^0], [61^0 - 80^0], [81^0 - 100^0], [101^0 - 120^0], [121^0 - 140^0], [141^0 - 160^0], [161^0-180^0]. Each bin will be calculated magnitude statistics at each location: at position (x, y) if the orientation belongs to that bin, the bin value at position(x,y) equals the magnitude value, otherwise the bin value at position(x,y) is 0.

In the next step, the algorithm computes the feature vector for each cell (each cell is usually selected with a size of 8 × 8 pixels). The feature vector consists of 9 elements that correspond to 9 bin; and the value at element i is the sum of values of points in bin i that have coordinates in that cell. Then, the feature vector of each block is calculated. Each block is usually chosen with a size of 2×2 cells (16 × 16 pixels). Vector features of blocks are calculated by combining feature vectors of each cell in the block together. The number of elements of the feature vector at each block is given by the formula:

$$Size_{feature/block} = n_{cell} \times Size_{feature/cell}$$

In which: $Size_{feature/block}$ is feature in a block; n_{cell} is the number of cell in a block; $Size_{feature/cell}$ is the number of feature in a cell.

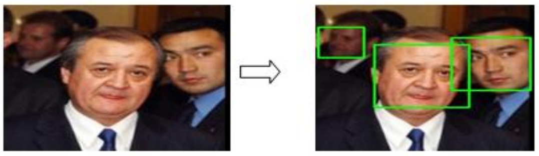
On the assumption that the size each cell is 8×8 pixels, the size of each block is 2×2 cells (16 × 16 pixels), the scale-orientation space is from 0^0 to 180^0 and is divided into 9 bins. Therefore, the number of feartures in each block is equal to 4×9 = 36 elements. From there, the algorithm computes the feature vector of windows over the entire input image, in which a window is overlapping blocks. The feature of a window is computed by pairing the feature vectors of each block that produces that window. The number of elements of each window is defined as follows:

$$n_{block/window} = \left(\frac{W_{window} - W_{block} \times W_{cell}}{W_{cell}} + 1 \right) \times$$

$$\times \left(\frac{H_{window} - H_{block} \times H_{cell}}{H_{cell}} + 1 \right)$$

$$Size_{feature/window} = n_{block/window} \times Size_{feature/block}$$

In which: W_{window}, $W_{Block,}$ W_{cell} are the width of window, block and cell (in pixel); H_{window}, $H_{Block,}$ H_{cell} are the height of window, block and cell (in pixel); $n_{Block/Window}$ is the number of block in a window, $Size_{Feature/Window}$ is the number of feature in a window.



(a) Khung hình đầu vào

(b) Các khuôn mặt được phát hiện

Figure 3. Detected face from input frame

In the final step, all obtained feature vectors on each window will be used as the input of the linear classifier SVM [12]. The classifier is responsible to determine the pattern (face or non-faceted) for each input image based on the knowledge that the algorithm has been trained on. Figure 4 - b shows the results of the human face detection algorithm on a specific input image.

B. *Face recognition*

The recognition process consists of two main steps: feature extraction and face classification. For the first step, we applied the model proposed by Florian Schroff et al. [5]: FaceNet neural network neural network. The advantage of the model is that the capability of learning from a given pattern set to automatically detect the most important features for object recognition.

This approach is based on learning a Euclidean embedding per image using a deep convolutional network. The network is trained such that the squared L2 distances in the embedding space directly correspond to face similarity: faces of the same person have small distances and faces of distinct people have large distances [5].

The network is directly trained to have a compact 128-D embedding using a triplet- based loss function. A triplet consists of two matching face thumbnails and a non-matching face thumbnail. The objective of loss function is to separate the positive pair from the negative by a distance margin.

From the obtained measures, the algorithm will estimate the value of the cost function by comparing the distance between two feature sets in which a set are generated from two different faces of a same person (the first person) and the other one are generated from the face of the other person (the second person). After estimation, the back-propagation is performed from the last layer to the first layer to refine the

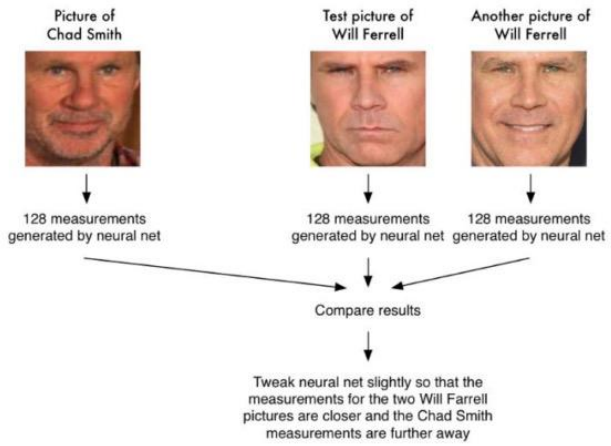

Picture of Chad Smith
Test picture of Will Ferrell
Another picture of Will Ferrell
128 measurements generated by neural net
128 measurements generated by neural net
128 measurements generated by neural net
Compare results
Tweak neural net slightly so that the measurements for the two Will Farrell pictures are closer and the Chad Smith measurements are further away

Figure 5. FaceNet

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

weights (weight updating). The computation, estimation and updating of network weights is repeated continuously until the value of the cost function satisfies the given condition. These steps are also repeated on overall training dataset until the training algorithm converges. The model is described in Figure 6.

Experimental results show that the deep neural network has a higher accuracy in feature extraction. Because the algorithm is trained with large, diverse data sets, therefore the detected features are less likely to be affected by the noise and the tilt and rotation properties of images. However, because the network is multi-layered architecture and the number of links between the network layer is very large, so the calculation on the network usually take a long time. So that, the overall speed of the algorithm is affected. Therefore, to ensure that algorithms can respond in real-time, we use the Graphic Processing Unit (GPU), which speed-up the calculations on network layer follow the parallel mechanism.

IV. EXPERIMENTS

We have realized experiments in order to show the efficient of the proposed method. The experiments are performed by using various open source/ or free libraries, such as NumPy libraries [24] for data representation, storage and manipulation, OpenCV library [23] for performing mechanical imaging tasks, Scikit-Learn library [25] for testing machine learning models (neural networks, svm models, etc.). The experiments is realized on Windows 10, with 2.4GHz processor, and 6GB RAM.

The proposed method is evaluated via standard database sets (containing frames taken from different camera devices, webcams) [26]. They are common databases for research groups. In this experiment we use:

- UOF dataset: Provided by the University of Essex, UK, consists of four data sets: faces94, faces95, faces96 and grimace. Pictures in the database are 24-bit color images in JPEG format. The data set contains a set of 395 personal pictures (both men and women) with 20 photos for each person, totaling 7900 photos. All faces are mainly performed by first-year undergraduate students ages 18 to 20 and some older adults, some individuals with glasses and beards, of different races.
- FEI dataset: The FEI face database is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory of FEI in Sao Bernardo do Campo, Sao Paulo, Brazil. There are 14 images for each of 200 individuals, a total of 2800 images. All images are colorful and taken against a white homogenous background in an upright frontal position with profile rotation of up to about 180 degrees. Scale might vary about 10% and the original size of each image is 640x480 pixels. All faces are mainly represented by students and staff at FEI, between 19 and 40 years old with distinct appearance, hairstyle, and adorns. The number of male and female subjects are exactly the same and equal to 100.

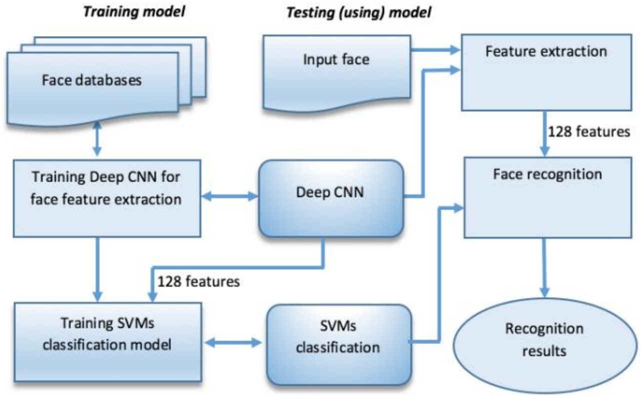


Figure 6. Face recognition

- JAFFE dataset: The dataset contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects.
- LFW dataset: the dataset contains 13,233 images with 5,749 identities, and is the standard benchmark for automatic face verification.

The evaluation process contains two stages: (i) evaluation of the face detection model and evaluation of the accuracy of face identification. The performance of the face recognition model is evaluated by the following measures:

- Detection Precision: **DP** = Number of correct identification over entire number of identification.
- Detection Recall: **DR** = Number of correct identification/ /(Number of correct identification + Number of unknown identification)
- Detection F-Measure:

DM = (2×FDP*FDR)/ (FDP+FDR)

Besides, we compared also the performance of the proposed face recognition model with this one using Haar-Like feature and AdaBoost classifier (Haar-Like AdaBoost) [19]. The experimental results are described in detail in TABLE 1.

The performance of the overall recognition model is evaluated by Precision metric as follows:

R_Precision = Number of correct recognized face/number of overall faces.

Table 1. Performance evaluation of proposed method

Data	Number	Proposed method			Haar-Like AdaBoost		
		DP	*DR*	*DM*	*DP*	*DR*	*DM*
Faces96	3040	98.42	99.2	98.81	93.75	94.50	94.12
FEI_P1	700	98.43	98.43	98.43	80.71	80.71	80.71
FEI_P2	700	99.14	99.14	99.14	83	83.00	83.00
FEI_P3	700	97.43	97.43	97.43	79.43	79.43	79.43
JAFFE	213	100	100	100	100	100	100
LFW	13233	99.74	99.74	99.74	93.27	93.27	93.27

The evaluation is performed consecutively on selected datasets. Each dataset was randomly divided into two set: training and testing (90% for training and 10% for testing).

The training consists of two stages: training the selective extractor (FaceNet) and training the SVM classifier (see Figure 6).

The experimental results are described in Table 2. In which, the of the proposed model is compared with the classification method using the PCA feature and the Eigenface classifier.

Table 2. Accuracy evaluation of proposed method

Data	Number of faces need to recognize	R_Precision (%)	
		Proposed methode	*PCA- Eigenface*
Faces96	3040	98.02	83.23
FEI_P1	700	98.16	82.12
FEI_P2	700	98.74	83.62
FEI_P3	700	97.55	75.43
JAFFE	213	99.02	95.17
LFW	13233	95.26	78.13

Based on the experimental results, the proposed method has high accuracy (over 95%) on all test data sets. Meanwhile, the accuracy of the PCA-Eigenface method is influenced by the brightness and movement of the input image.

V. CONCLUSION

In this paper, we propose an overall solution in face recognition from a live webcam. The proposed model contains two main step: face detection and face recognition. The perfomance of model has been evaluated on standard data sets, which are shared by the worldwide facial recognition research community, including UOF, FEI, JAFFE and LZW databases. The evaluation was divided into two steps, in which the perfomance of the face detection method was evaluated based on three levels of precision, recall and F-measure, while the efficiency of the face recognition model is evaluated based on the recognition accuracy. Experimental results show that the proposed model achieves high accuracy and stability.

REFERENCES

[1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[2] Davis E King. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10:1755–1758, 2009.

[3] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

[4] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, ¨Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. The Journal of Machine Learning Research,12:2825–2830, 2011.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.

[6] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[7] Hiyam Hatem, Zou Beiji,Raed Majeed, "A Survey of Feature Base Methods for Human Face Detection", International Journal of Control and Automation Vol.8, No.5 (2015), pp.61-78.

[8] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. IEEE International Conference on Image Processing (ICIP), 265(265):530, 2014.

[9] Hwai-Jung Hsu and Kuan-Ta Chen. Face recognition on drones: Issues and limitations. In Proceedings of the First Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use, DroNet ’15, pages 39–44, New York, NY, USA, 2015. ACM.

[10] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. JOSA A, 4(3):519–524, 1987.

[11] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection. IEEEcComputer Society Conference on Computer Vision and Pattern Recognition, 2005.

[12] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In Computer Vision, 2009 IEEE 12th International Conference on, pages 365–372. IEEE, 2009.

[13] Neeraj Singla, IISugandha Sharma, "Advanced Survey on Face Detection Techniques in Image Processing", International Journal of Advanced Research in Computer Science Technology (IJARCST 2014), vol. 2 Issue 1 Jan-March 2014.

[14] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. Proceedings of the British Machine Vision, 1(3):6, 2015.

[15] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. JIPS, 5(2):41–68, 2009.

[16] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. Neural Networks, IEEE Transactions on, 8(1):98–113, 1997.

[17] Turk, M. and Pentland , A. 1991. Eigenfaces for recognition. J. Cogn. Neurosci. 3, 72–86.

[18] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1867–1874, 2014.

[19] Viola, P. and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In Proceedings, IEEE Conference on Computer Vision and Pattern Recognition.

[20] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8.

[21] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 7, 8, 9.

[22] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonicalview faces in the wild with deep neural networks. CoRR, abs/1404.3543, 2014. 2

[23] http://opencv.org/

[24] http://www.numpy.org/

[25] http://scikit-learn.org/stable/

[26] http://www.face-rec.org/databases/

First Author: Dr. Thanh Tan Nguyen Thi was born in HaNoi , Vietnam, in 1977. She received the B.S. degree and M.S. degrees in information technology from VNU University of Engineering and Technology, in 2004 and the Ph.D. degree in computer science from Vietnam Academy of Science and Technology, in 2012. From 1999 to 2012, she was a Researcher at the Pattern recognition and knowledge discovery department, Institute of Information Technology, Vietnam Academy of Science and Technology. Since 2013, he has been the Head of of Computer science and Information Systems Department Information Technology Faculty - Electric Power University, in HaNoi, VietNam. She has extensive experience in the AI, Image Processing, Pattern Recognition including Optical Character Recognition, Form Recognition, Biometric Recognition fields in which she has been pulic more than 20 scientific papers. She is one of the authors of the VnDOCR (Vietnamese Optical Character Recogniton) software, MarkRead (check mark form auto reader) software, VnIDCard (Vietname Identify Card Auto Reader) sofware. Now, she is focusing on the proplem of developing the Security Camera Systems based on face recognition.

Second Author: Dr. Khanh Nguyen Trong was born in 1982. He received his Bachelor degree in Information Technology (IT) at Hanoi University of Science and Technology in 2005, his Master degree in IT at the L'Institut de la Francophonie pour l'Informatique (old IFI) in 2008, and his Ph.D. in IT at the University of Paris VI , France, in 2013. He is currently a lecturer at Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. His domains of interest are Distributed System, Computer Support Collaborative Work, Modelling and simulation of the complex system, Collaborative Simulation and Modelling.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.