EasyChair Preprint

№ 2183

An Effective Model for Intergrated Face Detection and Recognition

Huy Quang Tran, Nhat Tien Le, Quang Luong Nguyen, Dat Tan La and Thu Thi Anh Nguyen

December 16, 2019

AN EFFECTIVE MODEL FOR INTERGRATED FACE DETECTION AND RECOGNITION

Huy Quang Tran, Nhat Tien Le, Quang Luong Nguyen,
Dat Tan La, Thu Anh Thi Nguyen

Faculty of Advanced Science and Technology, The University of DaNang - University of Science and Technology, Danang, VietNam

quanghuypthiet@gmail.com ltn281097@gmail.com,
quang.cl.qh@gmail.com, ltdd117@gmail.com,
ntathu@dut.udn.vn.

Abstract. Facial recognition, an attractive field in computer-based application, has been one of the most widely research and challenging areas in computer vision and machine learning. The innovation of new face authentication technologies is a controversal topic to build much effective and robust face recognition algorithms. In this work, an effective, fast and reliable model is proposed based on combining traditional algorithms such as HOG, SVM and the modern ones such as ResNet50, Facial Landmark 68 for face recognition and emotion detection. Tests on different databases of large number of samples, various environmental conditions and facial expressions are presented with high recognition results.

Keywords: Histogram of Oriented Gradients (HOG), Residual Neural Network (ResNet), Support Vector Machine (SVM), Open Face Framework.

1 Introduction

Face recognition system can be applied in many practical aspects of life, such as security system, house unlock, phone unlock, tracking system, etc. There have been many challenges in reality that a face recognition system has to face with, such as high accuracy, short processing time, real-time response, robust recognition in difficult conditions (emotional expression, angle of view, illumination condition, to name a few).

Up to now, there have been many methods developed to classify and identify faces. Support Vector Machine (SVM) [1] and combination with independent component analysis [2] are typical techniques that show high recognition performance. Research reported in [3] indicates that the combination of SVM and Gabor filter is good for adapting changes of brightness, posture and facial expression. However, this method requires a large number of computation, so the processing speed is quite slow. Besides, the combination of Histogram of Oriented Gradient (HOG) method and SVM has been proved to be highly effective approach, according to another research [4]. A study reported in [5] shows that application of pre-processing techniques will lead to increase of recognition rate. Therefore, in this study, a face recognition method which bases on HOG method combining with SVM and ResNet50 has been proposed.

The whole model of face recognition and emotional detection including training phase and test phase is demonstrated in Figure 1.

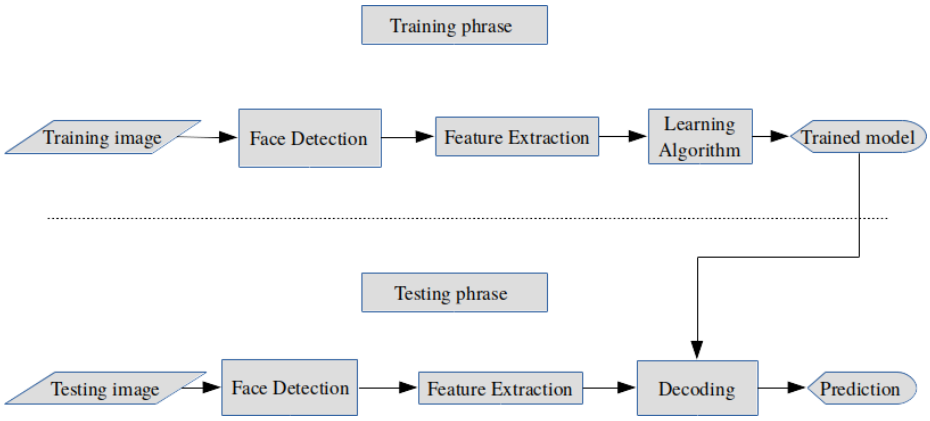

Fig. 1. The general diagram of the whole model.

Firstly, the input image is pre-processed for better feature extraction process with high effectiveness. Feature extraction step is then followed to extract unique features for face identities and emotional styles. In the next step, at the training phase, the image learning process will take place and create a new model. The trained model is named for later recall. In the testing phase, the decoding process bases on the trained model to conduct a prediction and eventually displays the name of the recognized face with the highest accuracy rate on the screenature extraction. The model is then evaluated through important criteria on different datasets for various facial postures, environmental conditions and large number of samples.

Details of the feature extraction method (HOG, SVM, ResNet50), training model, experiments, testing results, evaluation and conclusion are periodically presented and discussed in the following sections of this paper.

2 Face Detection and Feature Extraction

2.1 Histogram of Oriented Gradients (HOG) and Pattern Matching

In this research, Histogram of Oriented Gradients, a technique usually applied for object detection, and Pattern Matching are now utilized to detect human's face from the input image. Firstly, a feature vector is calculated using HOG on the input image. After that, a pre-trained SVM's model is used on many HOG's feature vectors of actual human face as the pattern. Each small blocks (of 16x16 pixel for example) of the image is compared to the pattern to check whether the block is human face. After searching through all the image, the size of the block can be resized in case the block is too small or too large in comparision to the size of the face in the image. If a match is found, the bounding box is saved for later using in the next session.

2.2 ResNet50 and Euclidean distance

For face recognition, a pre-trained RestNet50 is used on the face image from previous session to extract 128D vector which represents all the features of a face. ResNet50 is a configuration of the 50 layers Residual Network. In general, in a deep convolutional neural network, several layers are stacked and are trained to the task at hand. The network learns several low/mid/high level features at the end of its layers. In residual learning, instead of trying to learn some features, it tries to learn some residual.

Residual can be simply understood as subtraction of feature learned from input of that layer. ResNet does this using shortcut connections (directly connecting input of n^{th} layer to some $(n+x)^{th}$ layer. It has been proven that training this form of networks is easier than training simple deep convolutional neural networks and also the problem of degrading accuracy is resolved.

This ResNet is especially trained only for face recognition on a large face image dataset. The vector is used for both training phase and testing phase. In order to find out the similarity of two face images, the Euclidean distance is calculated as in Equation 1. It is a simple distance calculation but it is very effective on high dimentional vector. If the distance is small, two faces are more similar.

$$d(u,v) = \sqrt{\sum_{i=1}^{128} |u_i - v_i|^2} \quad (1)$$

3 Training Model

In the previous session, the pre-trained ResNet50 is built to extract 128D vector which represents all the features of a face. It contains encoded facial landmark of the training set and a parameter called threshold. The threshold is used to check the model's prediction. If the cost function value is smaller than the threshold, the prediction is accepted, whereas the prediction is marked as unknown. Because of that, the model only needs one typical face's image per individual.

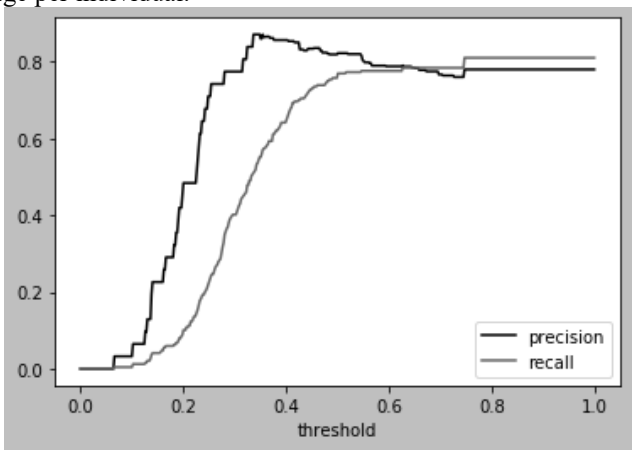

Fig. 2. Precision and recall per threshold.

We train the model on the training set time to time with different thresholds to find out the best result. Certainly, the best threshold on the training set is not the best threshold for applications but it does not vary too much. As the model performance is measured by precision and recall, we typically choose the point where the precision and recall are balanced, it is at the intersection of two lines as described in the Figure 2.

4 Evaluation

4.1 Criteria

4.1.1 Confusion matrix

Confusion matrix is a method to describe performance of a classification model. It is a table that the vertical axis is the true label and the horizontal axis is the predicted label. The cell of the table contains a number of how many times it falls into that case. The advantage of using confusion matrix is that we have an intuitive way to see how good our model performs per class. In this study, classification measures are used to examine performance of the proposed face recognition technique. Accuracy, precision, recall and f1-score are calculated from the confusion matrix through basic terms as follows:

- True positives (TP): is an outcome where the model correctly predicts the positive class.
- True negatives (TN): is an outcome where the model correctly predicts the negative class
- False positives (FP): is an outcome where the model incorrectly predicts the positive class
- False negatives (FN): is an outcome where the model incorrectly predicts the negative class

4.1.2 Accuracy

Accuracy describes how often the model predicts correctly. It is calculated by the following formula:

$$ACC = (TP + TN)/ (TP + TN + FP + FN) \quad (2)$$

4.1.3 Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$PR = TP / (TP + FP) \quad (3)$$

4.1.4 Recall

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$RE = TP / (TP + FN) \quad (4)$$

4.1.5 F1 score

This score is calculated from precision and recall, use this score when we seek the balance between two scores.

$$F1 = 2 * (PR*RE)/(PR+RE) \quad (5)$$

4.2 Databases and Scenarios

Three different face datasets named as Yale, AT&T, FERET are used for evaluating performance of the proposed face recognition technique. As the proposed technique uses only 1 training image sample, the mentioned datasets are reconstructed into two different senarios of training and test as follows:

- The Well-Matched (WM) test set is the set of images that look not too different from the ones in the training set. It may contain images with lower resolution, quite left or right aligned and less light condition variance.
- The Highly Miss-matched (HM) test set consists of images that show much difference between training and test set. It can contain completely different face poses, face position and light condition. The test set consists of similar faces to the training set but their background and the light condition do not match much to the ones of the training set.

4.2.1 AT&T Face Dataset

The AT&T dataset [6] is collected from 40 individuals, 10 images per person and contains many face poses with different emotions on the face. This dataset is used to check if the model work well on all case of face poses with emotions.

In this test, we reserve 5 classes (50 images) from the dataset and not put it in the train set to test the attack prevention ability of the model. After testing on 40 individuals with 9 images per person, average recognition rate is calculated from the confusion matrix. The HM scenario is applied for this dataset as there is significant difference between training and test data in terms of face poses and emotion.

4.2.2 FERET Dataset

The FERER dataset [7] consists of 14,126 images of 1199 individuals collected from the year of 1993 to year 1996, and contain 365 duplicate sets of images. A duplicate set is a second set of images of a person already in the database and was usually taken on a different day. For some individuals, over two years had elapsed between their first and last sittings, with some subjects being photographed multiple times. This time lapse was important because it enabled researchers to study, for the first time, changes in a subject's appearance that occur over a year.

This dataset is used to find out what happens when the number of classes grows enormous. Therefore, the WM scenario is applied only for this dataset.

4.2.3. The Yale Face Dataset

The Yale dataset [8] contains 165 gray-scale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised, and wink. This dataset is used to test the model's performance on different light condition and different facial expression.

There are 14 individuals selected for training and test. Each individual consists of 1 frontal face image from this dataset. One individuals is not included in the train set to check if the built model can exclude it from other classes.

With this dataset, the WM and HM secarios are both cosidered for the test. The WM test set consists of 4 images with the following attributes: wearing glasses, not wearing glasses, sad, sleepy. The HM test set consists of 6 images with the following attributes: happy, surprised, winky, light left, light right, center light.

Figure 3 shows some examples which present for each scenario.

WM test set example image:

Normal face (a)

Sad face (b)

HM test set example image:


Right light on face (c) Surprised face (d)

Fig. 3. Different face postures.

4.3 Results and Discussion

The results of the method is compared to the method in another article using HOG features with SVM classifier [11].

The obtained results show in Table 1 demonstrates that the proposed model performs very well on the Yale dataset. The accuracy scores of 100% are obtained for both WM and HM scenarios because the dataset doesn't have noise in the background, light condition is fairly good. Though the dateset contains both emotional faces and non-emontial faces, the model still recognizes them perfectly and the model doesn't miss-classify the unknown individuals to any class. With the FERET dataset, the recognition results show high accuracy and F1 scores. With the AT&T dataset, the recognition results show a drop down of about 15% of accuracy and F1 scores. A detailed analysis on derived confusion matrix reflects that the built model miss-classify target face class into other classes. This is because of huge difference between test images and training images.

The obtained results are compared to the published results in another research using HOG features with SVM classifier [11].

The proposed method give leads to a big performance improvement in comparing to traditional HOG features and SVM classifier method. It is assumed that The reason behind is the RNN, it which gives a more detail features vector, contributes thus our classifier is quite simple but the model to better performance recognizerperformance is great.

Table 1. Accuracy/ F1 Score (%) results (* not applied)

	Our method		HOG & SVM [11]
	WM	HM	Overall accuracy
1. AT&T Face Database	(*)	85.55 / 86.23	92.5
2. Color FERET Database	98.56 / 99.64	(*)	68.5
3. The Yale Database	100.00 / 100.00	100.00 / 100.00	92

The AT&T test set are also separated into 5 categories (Center, Half-left, Half-right, Head-up, Head-down poses) as expressed in Table 2 in order to test performance of the trained model on each face pose. After labeling, the dataset only has 28 individuals with sightly different face poses with emotions. The results show that the model is able to detect Center, Half-right and Head-up face poses. However, the Half-left and the Head-down face poses present lower performance. This is probably because of the differences between the Head-down and the Center poses as well as with facial emotions. The Head-up pose is quite similar to the Center one but the Half-left and the Half-right poses display difference in Accuracy. Figure 4 demonstrates the results in increasing trend of obtained Accuracy and F1-score with standard deviation.

Table 2. The result of testing with face poses.

Attribute	Center	Half-left	Half-right	Head-up	Head-down
Accuracy(%)	85.18	75.17	90.32	92.10	76.47
F1-score(%)	80.00	73.06	82.45	84.00	60.42

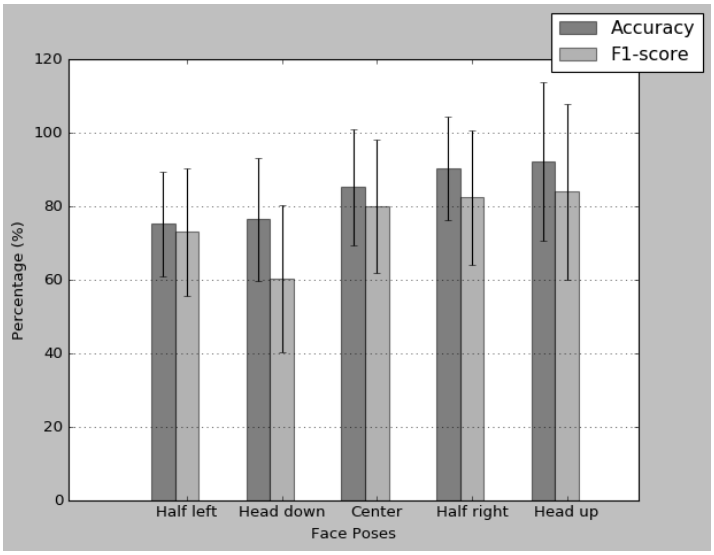

Fig. 4. The result of testing with face poses.

After testing on 40 individuals with 9 images every person, the general result is calculated based-on the confusion matrix. In this test, we reserve 5 classes (50 images) from the dataset and not put it in the train set to test the attack prevention ability of the model. The model sometimes miss-label that the sample is unknown because of huge different between test image and train image. The solution for that is to align face to frontal pose or adding that face pose to the train data. The results show that the model accuracy drop down about 10% on the head-down face pose. This is true because the model miss-classify those images into other class but the results is still acceptable.

5 Conclusion

In this study, we have introduced an effective model for face recognition combining facial emotional detection. The model utilizes traditional algorithms such as HOG, SVM and the modern ones such as ResNet50, Facial Landmark 68. The test results have shown that this method works well with different datasets: the AT&T Face Datase for images of faces at different postures, along with different illumination conditions and facial emotional expressions; the FERET Dataset for large data set with over 1000 different faces; the Yale Face Dataset for faces with facial emotional expressions. It also shows that the performance of this proposed algorithm is quite high and reliable even in different conditions such as various lighting, various camera views, and various face emotions. This model can be perfectly applied for many face's pose and many face's expression.

Acknowledgment

This work was supported by The University of Danang, University of Science and Technology, code number of Project: T2019-02-41. Specially thanks to Assoc.Prof.Dr. Tuan Van Pham, FAST, DUT, who provided insight and expertise that greatly assisted the research in this paper.

References

1. Juneja, Komal, An improvement on face recognition rate using local tetra patterns with support vector machine under varying illumination conditions, IEEE Computing. 2015 International Conference on Communication & Automation (ICCCA), India, pp. 1079 – 1084, May 2015.
2. Jia Jun Zhang, Yu Ting Shi, Face recognition systems based on independent component analysis and support vector machine, IEEE 2014 International Conference on Audio, Language and Image Processing (ICALIP), Shanghao, pp. 296 – 300, July 2014.
3. G. Majumder, M. K. Bhowmik, Gabor-Fast ICA Feature Extraction for Thermal Face Recognition Using Linear Kernel Support Vector Machine, IEEE 2015 International Conference on Computational Intelligence and Networks (CINE), Bhubaneshwar, pp.21-25, Jan. 2015.
4. H. S. Dadi, G. K. M. Pillutla, Improved Face Recognition Rate Using HOG Features and SVM Classifier, IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), Volume 11, Issue 4, Ver. I, pp. 34-44, Jul-Aug. 2016.
5. Nawaf Hazim Barnouti, Improve Face Recognition Rate Using Different Image Pre-Processing Techniques, American Journal of Engineering Research (AJER), Volume 5, Issue 4, pp. 46-53, 2016
6. AT&T Laboratories, The Database of Faces, April 1994
7. CambridgeP. J. Phillips, H. Wechsler, and P. Rauss. The FERET database and evaluation procedurefor face-recognitionalgorithms. Image and Vision Computing, 16(5):295–306, 1998.
8. Yale University. Yale Face Database. http://vision.ucsd.edu/content/yale-face-database, May 31, 2001
9. King, D. E. (2009). "Dlib-ml: A Machine Learning Toolkit" (PDF). J. Mach. Learn. Res. 10 (Jul): 1755–1758.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. “Deep residual learning for image recognition”. In: Conference on Computer Vision and Pattern Recognition, CVPR 2016, 2016, pp. 770–778
11. Dadi, H. and Mohan Pillutla, G. (2016). Improved Face Recognition Rate Using HOG Features and SVM Classifier. *IOSR Journal of Electronics and Communication Engineering*, 11(04), pp.34-44.