COMPARISONS OF FACE RECOGNITION METHODS

Kai Wey Lim
20011792
hfykl2@nottingham.ac.uk
School of Computer Science, University of Nottingham

ABSTRACT

The paper compares the performance between non-deep learning and deep learning approaches in facial recognition. Accurate classification and training time of the methods are used for comparison. A total of 3 face recognition methods are used in this paper, namely FaceRecognition, FaceRecognition1 and FaceRecognition2. FaceRecognition is the baseline method. FaceRecognition1 is the non-deep learning approach that uses Viola-Jones algorithm for face detection and Histogram of Oriented Gradients (HOG) for feature extraction. FaceRecognition2 is the deep-learning approach that is known as VGG-Face, a Convolutional Neural Network (CNN) that is trained on a large face dataset [8]. The face recognition methods are trained on the same dataset which consists of 100 face images that represents a unique identity each. The face recognition methods are then evaluated on its classification accuracy on the test data which consists of 1344 face images of identities from the training data. Results from experimentation have showed that the deep learning approach outperforms the other approaches.

1. INTRODUCTION

Face recognition has been one of the major topics discussed and researched in the computer vision field. Multiple algorithms have been developed to increase its accuracy that surpasses its predecessors. More recent methods which are deep learning methods are known to be more robust in its recognition task with challenges such as pose, light illumination, occlusions, and misalignments [1]. Face recognition algorithms that will be discussed in this paper can be separated into deep learning approach and non-deep learning approach.

Non-deep learning approach would be traditional face recognition methods that uses feature descriptors such as Histograms of Oriented Gradients (HOG), Local Binary Pattern (LBP), and Scale Invariant Feature Transform (SIFT) to extract features from the face in the image, where the features would then be used to train machine learning algorithms such as Support Vector Machine (SVM) for classification [1]. Previous studies have shown that the use of HOG features for face recognition have yielded promising results, where they experimented on the impact of facial landmark localization and the impact of HOG features extraction from grid with multiple scales [2].

Deep learning approach consists of Convolution Neural Network (CNN) methods. CNNs are comparable to traditional Artificial Neural Network in which both contains neurons that self-optimizes through learning [3]. CNNs take images as input data and carry out processes such as convolving images with filters, and pooling to down-sample images for complexity reduction in the network's deeper layers, which allows the network to discover different representations of features or patterns in the input image [1, 4]. The feature encodings extracted from the previous layers will then be passed into the fully connected layers, which produces class scores that is used for classification in the end layers of the network [3]. The final layer of the network normally consists of loss functions such as Softmax, which is widely used in CNNs due to its simple and clear probabilistic interpretation for training the model to classify input images [5]. Developed CNNs are known for their achievement in state-of-the-art accuracies but requires a significant amount of data and time to train [1, 8, 9].

This paper will be discussing about the comparison between three facial recognition algorithms, with each from one of the approaches mentioned and one baseline method. All the approaches will be trained on 100 unique faces with a single training data for each face. The algorithms will then be tested on a total of 1344 images that consists of different images that represents the faces that the algorithms have trained on. Deep learning methods are often known for its high accuracy and high training time while non deep learning usually produce lower accuracy results and requires lower computational time. The hypothesis in this discussion will be that deep learning methods for face recognition will outperform the non-deep learning methods.

The discussions for the remaining sections in the paper are as follows. Section 2 describes the face recognition methods that are used for comparison in the paper. Section 3 portrays and reviews the results obtained by the face recognition methods. Section 4 discusses on the performance of the methods. Section 5 concludes the paper.

2. FACE RECOGNITION METHODS

Algorithms and structures of all 3 different face recognition methods used will be discussed in this section.

2.1. Face Recognition

A baseline method is introduced to compare between the two face recognition approaches. The baseline method turns the input images to grayscale and then normalizes its intensity. Zero mean normalization is then performed on the images to be used as feature vectors. The feature vectors are accumulated for all the faces in the training data and compared with the faces in the testing data through cross-correlation based template matching. The testing images are classified as the training data's label that produces the largest corresponding value from cross-correlation.

2.2. Face Recognition 1

The first face recognition method uses the non-deep learning approach. The input images are first turned into grayscale images to negate the use of colored images, which reduces the range of the pixel values. The images are also resized to 300x300 and every pixel is turned into a double precision ranged from 0 to 1, to reduce computational time. Based on the average time collected from running the algorithm for 3 trials, it has shown that changing the images into double precision has reduced the algorithm's runtime by 85.44 seconds.

In this face recognition method, Viola-Jones algorithm is introduced to rapidly detect the faces in each input image to discard the irrelevant parts of the image that is not the face. There are three key features used in the Viola-Jones algorithm which are, integral images, AdaBoost, and attentional cascade that constructs the face detection system [6]. Integral images are used for fast feature evaluation, whereby rectangle features are computed rapidly with an intermediate representation of the image [6]. A variant of AdaBoost is used to select features and train the classifier, where the learning algorithm is used to boost the classification performance of a simple learning algorithm [6]. The attentional cascade, an algorithm that constructs a cascade of classifiers that improves the performance of detection and significantly reduces computational time [6]. In this face recognition method detected faces are cropped from the image with the size of 165 x 165, while the images where faces could not be detected retain its content and resized into 165 x 165 to retain the consistency.

With the use of Viola-Jones algorithm, it allows the feature extraction to concentrate on the face instead of irrelevant parts of the image. The following step is the extraction of facial features with the image descriptor, Histogram of Oriented Gradients (HOG). HOG characterizes the distribution of local intensity gradients or edge directions of local object appearances, without the precise knowledge of the corresponding edge positions [7]. Each image is divided into cells for the accumulation of gradient directions or edge orientations that is turned into a local 1-D histogram, the combined histogram of each image then forms the representation [7]. Through this, the general shape and features of an individual's face can be extracted for training. Zero mean normalization is also applied on the extracted features from the HOG function. By applying zero mean normalization it has increased the method's accuracy by an average of 12% as well as reduced its runtime by an average of 388.66 seconds.

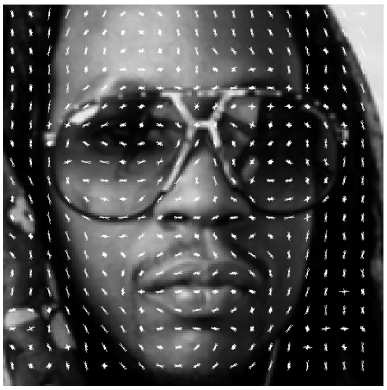

Figure 1: HOG feature representation of a face plotted over the face cropped from Viola-Jones algorithm.

The obtained features for every individual's face are used to train the supervised machine learning algorithm, Support Vector Machine (SVM). For faster computational time, the SVM is set to be trained as a linear classification model with the observations set corresponding to rows. Experimental trials from running the algorithm have shown that setting the SVM as a linear classification model have reduced the runtime by 964.01 seconds.

2.3. Face Recognition 2

The second face recognition method uses the deep learning approach. Transfer learning of a pre-trained CNN model, VGG-Face is used in this face recognition method. VGG-Face is inspired by the CNN model, VGG-16 trained on a very large-scale image dataset of faces that is used to extract face representations [8]. VGG-16 is a CNN model that has achieved state-of-the-art accuracy, which takes 224x224 colored images as input [9]. The architecture of VGG-16 is made up of a stack of convolutional layers that uses 3x3 filters that has convolutional stride of 1 pixel, such that the spatial resolution is preserved after convolution [9]. Additionally, VGG-16 have 16 adjustable weight layers and 138 million parameters [9]. The following layers after the convolutional layers are three Fully Connected layers which is followed by a Softmax layer for classification [9]. The notable difference between VGG-16 and VGG-Face is that the final fully connected layers in VGG-16 are known as "convolution" for their special case of convolution, because the filter sizes and the input data sizes are the same, hence each filter "senses" data from the whole image [8].

As stated, there is only 1 training image for every individual's face in the training dataset. Due to the lack of training data to train a deep neural network, this face

recognition method will import the VGG-Face model and have its weights freeze up until the 15th adjustable weight layer. The final layer with learnable weights will be adjusted during training, along with the replacement of the Softmax activation function to learn the classification of the new classes in the dataset. This allows the network to retain most of its pre-trained weights, which retains the learned weights for facial feature extraction. By doing so, the training time to retrain the network will be greatly reduced as there is a significant reduction of weights required to be adjusted. At the same time, the 16th layer's adjustable weights and new Softmax layer allows the network to relearn its classification to classify the new faces in the training dataset.

After the replacement of the Softmax activation function and freezing of weights, the whole network is then retrained on the training dataset so that it can classify the new faces. A total of 6 trials have been tested on the algorithm which freezes the 9th adjustable weight layer until the 16th adjustable weight layer. The average results of these trials have shown that freezing the 15th adjustable weight layer produced the best result with the lowest average runtime of 372.8 seconds while the freezing of the 10th adjustable weight layer produced the highest average runtime of 799.06 seconds. The accuracy of the method from the freezing of the different adjustable weight layers only differed from each other by a maximum of 6% accuracy.

3. RESULTS

With the finalized algorithm and parameters set for each face recognition method, the methods are then used to compare against each other. Each face recognition method is trained on the training dataset. For a fair comparison between the methods, they are trained purely on the 100 images without any additional augmented images and tested on the same test set. The comparison is then carried out by evaluating all the face recognition method's performance at accurately classifying individuals in the test data and the computational time it took for its training process. Each face recognition method returns a sequence of their label prediction corresponding to the test data passed through the algorithm. The prediction sequence is then used to compare against the ground truth label of the test data, to obtain the count of the correctly predicted test images by the face recognition methods. The count is then divided by the total size of the test data and multiplied by 100 to obtain the face recognition method's accuracy in percentage. A time counter in seconds, was used to time the training process of the face recognition method. Each face recognition method has been evaluated for 6 trials and their corresponding result is recorded. Table 1 shows the collected results of each face recognition method and their corresponding average result across the 6 trials.

	Face Recognition Methods					
	FaceRecognition		FaceRecognition 1		FaceRecognition2	
Trials	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)
1	55.625	25.372	179.869	38.244	318.056	81.3988
2	57.1462	25.372	178.486	38.3185	324.483	80.5804
3	61.1978	25.372	184.082	38.244	322.055	78.3482
4	59.1836	25.372	169.449	38.244	324.607	79.4643
5	56.9016	25.372	172.017	38.3185	321.954	80.5804
6	63.2496	25.372	182.456	38.244	333.048	80.0595
Average	58.884	25.372	177.727	38.268833	324.034	80.071933

Table 1: Comparison in accuracy (%) and time (s) between all the face recognition methods for 6 trials.

According to the collected results in Table 1, FaceRecognition2 produces the highest accuracy among all the other methods with an average accuracy of 80.07%, followed by FaceRecognition1 with an average accuracy of 38.27% accuracy and the lowest is FaceRecognition with an accuracy of 25.37% accuracy. In contrast of the accuracy results, FaceRecognition2 took the highest training time with an average of 324.03 seconds, followed by FaceRecognition1 with an average training time of 177.73 seconds and lastly FaceRecognition with 58.89 seconds.

To compare between the accuracies obtained the Wilcoxon Rank Sum test is introduced through the built-in function in MATLAB, ranksum(x,y). Wilcoxon Rank Sum test is a non-parametric test for the comparison between two populated samples, where the rejection of the null hypothesis indicates that the medians of the two samples are different from each other [10]. The results obtained from comparing the accuracy sample between FaceRecognition1 and FaceRecognition2 are $p = 0.0022$, $h = 1$ and rank sum = 21. Whereas the results obtained from comparing the accuracy sample between FaceRecognition and FaceRecognition1 are $p = 0.0022$, $h = 1$ and rank sum = 57. Both the comparison results prove the rejection of the null hypothesis of equal medians between the accuracies, at the default of 5% significance level. This indicates that the median of the accuracy performance between the face recognition methods are different from each other. Therefore, in terms of accuracy it can be stated that, FaceRecognition2 performed significantly better than FaceRecognition1, while FaceRecognition1 performed significantly better than FaceRecognition.

As for the comparison between the training time of the face recognition networks, it can be stated that FaceRecognition2 took 82.32% more time compared to FaceRecognition1. Whereas, FaceRecognition1 has an increase of 201.83% in training time compared to FaceRecognition. With this it can be stated that FaceRecogition2 took significantly more time to train compared to FaceRecognition.

4. DISCUSSION

From the significant increase in accuracy performance of FaceRecognition2 compared to FaceRecognition1 and

FaceRecognition, it can be perceived that FaceRecognition2 has a better performance in terms of accurately classifying faces because the pre-trained network was trained on facial feature extractions, hence allowing it to be more robust in classifying the faces with different challenges such as head poses, occlusions, blurring, and noises compared to FaceRecognition1. To further understand the classification performances of the face recognition methods, inspection of the performances is carried out by comparing and understanding the classified images.

Figure 2: Sample of misclassified faces by FaceRecognitionl where each row represents each run of classification. First column are the images that are getting predicted, second column are the images that the algorithm predicted, and the third column are the images from the training dataset which are the correct face corresponding to the face that is getting predicted.

It can be seen from the misclassified faces that FaceRecognition1 tend to classify faces that have the same head pose regardless of the difference in facial feature. Hence, it can be concluded the algorithm tends to be biased towards the pose of an individual's head and the facial features of an individual such as eyes and noses are not greatly affecting the classification process. From this, it can be interpreted that the feature extraction method used in FaceRecognition1 was not strong enough in extracting good and useful facial features to sway the classification process.

From the correctly classified faces from the test data by FaceRecognition2, it is seen that the algorithm is less prone to misclassify images from biased challenges like head pose. The method is also more robust towards challenges like noises, blur, and occlusions. Instead, the algorithm is more prone to misclassify images due to a more dominant feature on the individual's face such as the distinctive shape and size of the nose or the shape of hairstyle of the individual, that was not captured well in the training images because of angle or occlusion.

It is also recorded that FaceRecognition2 took significantly more time to train compared to the other methods. From the average training time of FaceRecognition1 it can be noted that setting the SVM as a linear classification model allows it to take lesser time to classify because it has lesser parameters to optimize. In comparison, FaceRecognition2 took longer to train because it has more parameters to optimize, such as the 16th adjustable weight layer and the optimization of the softmax activation function.

Figure 3: Sample of correctly classified faces done by FaceRecognition2 where each row represents each run of classification. First and third column are the images that are getting predicted. Second and forth column are the correct faces corresponding to the faces that is getting predicted.

Figure 4: Sample of misclassified faces by FaceRecognition2 where each row represents each run of classification. First column are the images that are getting predicted, second column are the images that the algorithm predicted, and the third column are the images from the training dataset which are the correct face corresponding to the face that is getting predicted.

5. CONCLUSION

From this paper, it can be concluded that in terms of classification accuracy deep learning method in this case outperformed the non-deep learning methods. Even though it took a longer time to train the deep learning method, the significant increase in accuracy outweighs the increase in training time. The robustness in feature extraction from the deep learning method allowed it to perform well. Hence, for tasks with small training data, transfer learning of pre-trained network is recommended to achieve high classification accuracy with a reasonable training time. On the other hand, for non-deep learning methods to outperform deep learning methods, a more robust feature extraction method would be required to strongly represent the face.

11. REFERENCES

[1] Mehdipour Ghazi, M., & Kemal Ekenel, H. (2016). A comprehensive analysis of deep learning based representation for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34-41).

[2] Déniz, O., Bueno, G., Salido, J., & De la Torre, F. (2011). Face recognition using histograms of oriented gradients. *Pattern recognition letters*, *32*(12), 1598-1603.

[3] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

[4] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1-6). IEEE.

[5] Wang, X., Zhang, S., Lei, Z., Liu, S., Guo, X., & Li, S. Z. (2018). Ensemble soft-margin softmax loss for image classification. *arXiv preprint arXiv:1805.03922*.

[6] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). IEEE.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[8] Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

[9] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[10] Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.