University Classroom Attendance System Using FaceNet and Support Vector Machine

Thida Nyein, Aung Nway Oo
University of Information Technology, Yangon, Myanmar
thidanyein@uit.edu.mm, aungnwayoo@uit.edu.mm

Abstract

Nowadays, face recognition system becomes popular in research area. Face recognition is also used in many application areas such as attendance management system, people tracking system, and access control system. For multi-face recognition, it has still many challenges for detection and recognition because it is not easy to detect multiple faces from one frame and it is also difficult to recognize the faces with poor resolution. Therefore, the main objective of this paper is to get a better accuracy for multi-face recognition by using the combination of FaceNet and Support Vector Machine (SVM). In this proposed system, FaceNet is used for feature extraction by embedding 128 dimensions per face and SVM is used to classify the given training data with the extracted feature of FaceNet. University Classroom Attendance System is applied by the proposed multi-face recognition. The Experimental result show that the proposed approach is good enough for multi-face recognition with an accuracy of 99.6%. It is better than VGG16 model on the same data-set.

Keywords- face recognition, deep learning, convolutional neural network, FaceNet, Support Vector Machine,VGG16.

1. Introduction

Nowadays, human face recognition is an important and popular technology used in many applications such as payment by using face recognition, unlock by face recognition, video monitor system, etc. Traditional attendance taking method is recording the attendance of students in sheet which takes a lot of time. Because of that system like automatic attendance is used to overcome the problems: consuming of time, incorrect attendance. This system is a computerized system for taking the attendance of students by using face recognition technology. There are three basic parts for face recognition technology (face detection/ pre-processing, feature extraction, and feature matching/classification). Deep Learning is used for face recognition because it is the best for recognition. For this proposed system, FaceNet is used for feature extraction and support vector machine is used for classification.

A face recognition system is a technology that can do identification and verification people from digital images and video surveillance. Nowadays, face recognition becomes trending in Artificial Intelligent fields. Face recognition is used in many applications to trace for crimes, to do payment, to do access right and to take attendance because it is reliable, inexpensive, and easy to use.

In the early 1990s, traditional face recognition systems were not stable and still not appear deep learning and have several errors in real time applications. Nowadays, deep learning appears and it is especially good for recognition and detection. Deep learning acts like a human brain, learning by itself. In deep neural network architecture, when we create a neural network, the more the hidden layers (neurons), the better the accuracy.

In this paper, FaceNet is used for extracting features from faces. Support vector machine is used for classification. To train the model, we used 'adam' optimizer and 'triplet_loss' as loss function. The purpose of the system is easily to know which students are in the classroom in a short time by implementing automated attendance management system by using face recognition technique and to save time consuming for taking attendance. The main purpose is to develop the reliable system by using deep learning. The system will record the attendance of the student automatically by matching from the training data-set.

2.ⁿ Related Work

For the past two decades, the research area is mostly on face detection [1]. Face detection has been an active research area and there are applied by using many traditional and deep learning methods [5]. In these days, both of detection and recognition are recently trending for research area. For face detection, many methods can provide for face detection with good accuracy, it means that there are almost completely perfected in face detection field. Therefore, it is more challenges for recognition. From early 1990s to near 2000s, holistic

learning approach (eigenface, fisherface, SRC and CRC,etc) and local handcrafted approach (LBP, HD-LBP, etc) were used for face recognition area respectively. Later 2010, shallow learning and deep learning become popular and nowadays, deep learning is the best for face recognition [8]. FaceNet [9] is a deep convolutional neural network. It provides achieves a new record accuracy of 99.63% and provides 95.12% on YouTube Faces DB [9]. In proposed system, FaceNet is used as feature extraction and it performs by embedding the features to 128 dimensions. After feature extraction, support vector machine is used as a classifier. Support vector machine (SVM)s [12] are a better learning algorithm than the Multi-Layer Perceptron (MLP) Classification for face recognition.

3. Proposed Approach

3.1 System flow

The flow of the proposed system is described in the following figure.

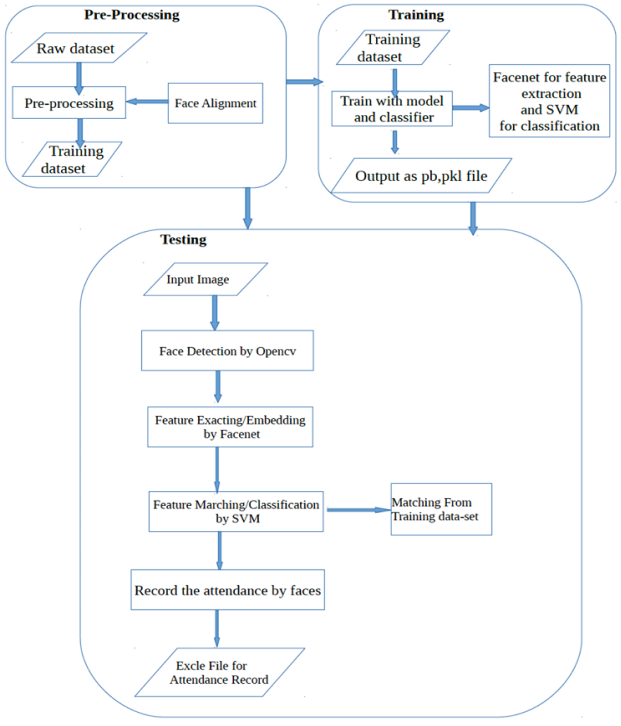

Figure 1. System flow of proposed system

For face recognition, we have to do three steps (preprocessing, feature extraction, classification). By using python libraries (PIL and face_recognition), the raw dataset of students' faces are pre-processed. FaceNet model is used for feature extraction. Features are extracted from pre-processed images. For classification (feature matching), support vector machine is used.

3.1.1 Pre-processing

Preprocessing is the important part of the proposed system. Train set and test set are pre-processed to detect faces from image by using PIL and face recognition libraries. The input image is defined as a numpy array for detection. When loading the image for detection, we can define the color channel with the mode that is format for converting the image to only 'RGB' (8-bit RGB, 3 channels) and 'L' (black and white).

For detection the faces, we can define how many times for sampling and can define the model. Even though "hog" model is faster on CPUs, "cnn" model is more accurate deep learning model. To create dataset for training, we need to resize (160x160x3) for each of the face images because the input shape is (160x160x3)for the training model(FaceNet).

Figure 2. Pre-Processing for training data-set

3.1.2 Feature extraction

FaceNet is built on the Inception Resnet architecture and there are generally 22layers. The weight of FaceNet is optimized using the triplet loss function, so that it learns to embed facial images into a 128-dimensions.

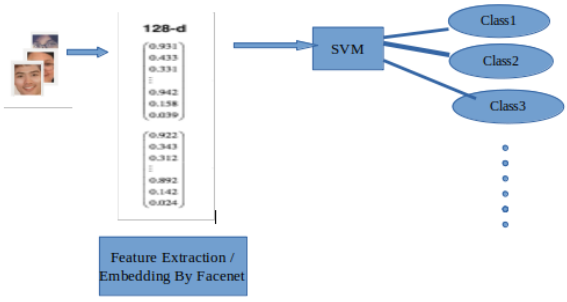

Figure 3. Steps of face recognition by FaceNet and SVM

When image classifications take an input image, computers see an input image as an array of pixels and also depend on the image resolution. For example, An

image of 6 x 6 x 3 array of matrix of RGB (3 refers to RGB values) and an image of 4 x 4 x 1 array of matrix of gray scale image. In this proposed system, 160x160 pixels images are input for model. In FaceNet, triplet loss is used as a special function among other loss functions. The Triplet Loss decreases the distance between an anchor and a positive input when both of which have the same identity, and increases the distance between the anchor and a negative input for the different identities.

Triplet Loss Function can be described like as

$$L(A,P,N) = \max(\|f(A)-f(P)\|^2 - \|f(A)-f(N)\|^2 + \alpha, 0)$$

where A is an *anchor*,
P is a *positive input, means that data from* the same class as A ,
N is a *negative input , means that data from the* different class from A,
α is a margin between positive and negative and
f is an embedding.

Steps of convolution neural network are convolution, pooling/sampling, activation/normalization, fully connected layers. There are three elements (input image, feature detector, feature map) for convolution operation. Input image is represented as an array. For input layer,a 5×5 or a 7×7 matrix is often used as a feature detector, but the more conventional one is a 3×3 matrix. The feature detector is also called as a "kernel" or a "filter".

Pooling is much like a filter for applying the feature maps. The size of the pooling operation or filter is smaller than the size of the feature map; specifically, it is almost always 2×2 pixels applied with a stride of 2 pixels. Therefore, the pooling layer will always reduce the size of each feature map by a factor of 2, that means, each dimension is reduced by half. For example, a pooling layer applied to a feature map of 6×6 (36 pixels) will result in an output pooled feature map of 3×3 (9 pixels). There are two types of pooling operation,max pooling and average pooling. Average pooling calculate the average value for each patch on feature map. Max pooling calculate the maximum value of each patch of feature map.

The activation layer controls how the signal flows from one layer to the next. Output signals which strongly depends on past references(past neuron result) would be enabling signals for propagation more efficiently for identification. For activation layer, Relu is mostly used and in this proposed system, Relu function is used for activation.

Table 1 -Architecture of CNN Model

Layer	Input Size	Output Size	Kernel	Strides
Conv1	160x160x3	160x160x64	3x3	2
Batch-normalization	160x160x64	160x160x64	-	-
MaxPooling	160x160x64	80x80x64	3x3	-
Conv2a	80x80x64	80x80x128	5x5	1
Batch-normalization	80x80x128	80x80x128	-	-
MaxPooling	80x80x128	40x40x128	5x5	-
Conv3a	40x40x128	40x40x128	5x5	1
Batch-normalization	40x40x128	40x40x128	-	1
MaxPooling	40x40x128	20x20x128	5x5	1
Conv4a	20x20x128	20x20x128	5x5	2
Batch-normalization	20x20x128	20x20x128	-	1
MaxPooling	20x20x128	10x10x256	5x5	1
Conv5a	10x10x256	10x10x512	3x3	1
Batch-normalization	10x10x512	10x10x512	-	1
MaxPooling	10x10x512	5x5x512	3x3	2
Conv6a	5x5x512	5x5x512	3x3	
Batch-normalization	5x5x512	5x5x512	-	
MaxPooling	5x5x512	2x2x512	3x3	
FC1	2x2x512	2x2x2048		
FC2	2x2x256	2x2x256		
FC512	2x2x256	1x1x512		

Table 2. Architecture of FaceNet

Layer	Input Size	Output Size	Kernel	Strides
Conv1	160x160x3	80x80x64	7x7x3	2
Pool1	80x80x64	40x40x64	3x3x64	2
Normalization	40x40x64	40x40x64		
Conv2a	40x40x64	40x40x64	1x1x64	1
Conv2	40x40x64	40x40x192	3x3x64	
Normalization	40x40x192	40x40x192		
Pool2	40x40x192	20x20x192	3x3x192	2
Conv3a	20x20x192	20x20x192	1x1x192	1
Conv3	20x20x192	20x20x192	3x3x192	1
Pool3	20x20x384	10x10x384	3x3x384	2
Conv4a	10x10x384	10x10x384	1x1x384	1
Conv4	10x10x384	10x10x384	3x3x384	1
Conv5a	10x10x256	10x10x256	1x1x256	1
Conv5	10x10x256	10x10x256	3x3x256	1
Conv6a	10x10x256	10x10x256	1x1x256	1
Conv6	10x10x256	10x10x256	3x3x256	1
Pool4	10x10x256	5x5x256	3x3x256	2
fc1	5x5x256	1x32x128		
fc2	1x32x128	1x32x128		
fc128	1x32x128	1x1x128		

Table 3. Architecture of VGG16

Layer	Input Size	Output Size	Kernel	Strides
Conv1	224x224x3	224x22x64	3x3	-
Conv2a	224x22x64	224x22x64	3x3	1
MaxPooling	224x22x64	112x112x64	3x3	1
Conv2	112x112x64	112x112x128	3x3	2
MaxPooling2	112x112x128	56x56x128	3x3	1
Conv3a	56x56x128	56x56x256	3x3	2
Conv3	56x56x256	56x56x256	3x3	1
MaxPooling3	56x56x256	56x56x256	3x3	1
Conv4a	56x56x256	228x28x512	3x3	2
Conv4	28x28x512	28x28x512	3x3	1
MaxPooling5	28x28x512	28x28x512	3x3	2
Conv5a	28x28x512	14x14x512	3x3	1
Conv5	14x14x512	14x14x512	3x3	1
FC	14x14x512	7x7x512	-	-
FC	7x7x512	7x7x512	-	-
FC4096	7x7x512	1x1x4096	-	-

3.1.3 Classification (Support Vector Machine)

Linear SVM is the extremely fast machine learning (data mining) algorithm for multiclass classification problems from large data sets. LinearSVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set[7]. The classification by SVM separate the classes from the extracted feature by FaceNet by calculating the distance. SVM algorithm implemented by scikit-learn is fast and easy to perform classification task.

4. Experiment

4.1 Data-sets

We used private face data-sets for training and testing. Face images are collected from my classmates and social media. We practically took some photos of celebrities from social media. Data-set is mixed with celebrities' photos. Total is over 80 persons. It includes about 10photos and some are 20photos for each. The input size of image is 160x160 pixels. The parameters used for training are initialized with the learning rate of 0.1, training with 1000 epochs. To train the model, we cropped and aligned the image and resized to 160x160, then using face detection technology to pre-process the faces (as shown in Figure 5) from original face images. For training data-set, the face images of each person are put each folder. The folder name will be label of person.

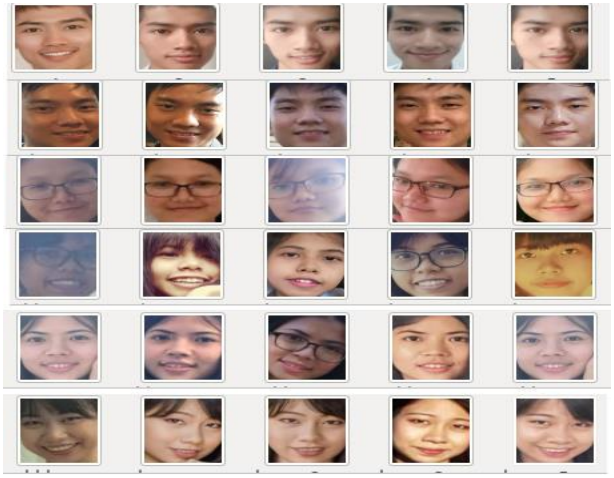

Figure 4. Example of Dataset

4.2 Experimental settings and result

The proposed system uses the face recognition algorithm from python libraries to detect faces. Therefore, it is mainly to install these libraries (Tensorflow (1.4.0), Scipy (0.17.0), Scikit-learn (0.19.1) and Opencv (2.4.9.1)). However, it takes more time when the dataset is big.

Using FaceNet and SVM, the accuracy is around 99.6 on testing. When using the general CNNmodel, it also gets around 95 accuracy and using pre-trained VGG16 got accuracy nearly 97. But the training time is different. FaceNet and SVM takes a little time but CNNmodel takes too long time. FaceNet and SVM work functionally and the batch size is 1000 for training model. CNN model is given that batch size is 20 and target size of face image is (160,160) for both training generator and validation generator.

Figure 5. Result from FaceNet and SVM

When training on CNN model, the number of epochs is 25 and samples_per_epoch is 8,000 and nb_val_samples is 2,000. Therefore, it takes too long for the training of model and can give with great accuracy. We use Adam as optimization algorithm in both CNNmodel and FaceNet. Moreover, categorical cross-entropy as loss function when the CNN model is compiled and triple loss is used in FaceNet, In CNNmodel, it includes fully connected layer and it works as classifier. After feature extraction by FaceNet, SVM performs for classification.

This proposed system gets multi-face recognition result with great accuracy by using FaceNet,and SVM. For face detection, Keras is used. Taking attendance in the classroom by face recognition is multi-face recognition. Therefore, the images of faces are poor resolution because this is not one by one recognition. It may be worse for detection when there are many people in one frame.. For face detection, we should use the deep learning technique. The maximum number of persons in the classroom is 20 when it is tested. There are still great for detection and recognition for 20 persons in one frame. It is easy to train the datasets with good accuracy for face recognition by using FaceNet and SVM. The feature extraction by FaceNet is only 128 dimension per face and that result is classified by SVM. The feature extraction by VGG16 is 4096 and that result is classified by fully-connected layer with Softmax function. It takes a little more time to train on training data-sets because VGG16 gives 4098 features for classification. Therefore, using FaceNet and SVM is easier to train than VGG16 and performance on testing data-set is also achieved with great accuracy. Therefore, using FaceNet and SVM is easier to train than VGG16.

Number of test image = TN+TP+FN+FP
Correct image number = TN + TP

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct)/(Number of all)

Table-4 The overall accuracy of FaceNet and SVM

Facenet and SVM (overall accuracy – 99.55)

Number of people	Number of train image	TN+TP+ FN+FP	TN+ TP	Accuracy
19	190	50	50	100
35	380	100	100	100
55	610	150	148	98.66

Table 5. The overall accuracy of VGG16

VGG16 (overall accuracy – 97.22)

Number of people	Number of train image	TN+TP+ FN+FP	TN+ TP	Accuracy
19	190	50	50	100
35	380	100	97	97
55	610	150	142	94.66

Table 6. The overall accuracy of CNNModel

CNNModel (overall accuracy – 95.43)

Number of people	Number of train image	TN+TP+ FN+FP	TN+ TP	Accuracy
19	190	50	49	98
35	380	100	95	95
55	610	150	140	93.33

5. CONCLUSIONS

This proposed system achieves the great accuracy for multiple face recognition when FaceNet is used as a feature extractor and SVM is used as a classifier. This proposed system aims to get reliable system by using multi-face recognition and can replace a manual system with an automated system. It will save time, reduce the amount of paper-work the administration has to do.

6. ACKNOWLEDGMENT

I would like to express my gratitude to all of my teachers for their advice. I am thankful for all people who support for this thesis.

7. References

[1] Sachin Sudhakar Farfade, Mohammad Saberian, Li-Jia Li,"Multi-view Face Detection Using Deep Convolutional Neural Networks",2015

[2] Syed Ibrahim, "Automatic Attendance System Using Facial Detection and Recognition Technique",2015

[3] Zhiyun Xue, Sameer Antani, L. Rodney Long, Dina Demner-Fushman, George R. Thoma,"Improving Face Image Extraction by Using Deep Learning Technique",2016

[4] Khem Puthea, Rudy Hartanto and Risanuri Hidayat,"A Review Paper on Attendance Marking System based on Face Recognition",2017

[5] Yicheng An, Jiafu Wu, Chang Yue,"CNNs for Face Detection and Recognition",2017

[6] Patrik KAMENCAY, Miroslav BENCO,Tomas MIZDOS, Roman RADIL,"A New Method for Face Recognition Using Convolutional Neural Network",2017

[7] Dr. Priya Gupta, Nidhi Saxena, Meetika Sharma, Jagriti Tripathi,"Deep Neural Network for Human Face Recognition",2018

[8] Mei Wang, Weihong Deng, "Deep Face Recognition: A Survey",2019

[9] Florian Schroff, Dmitry Kalenichenko , James Philbin "FaceNet: A Unified Embedding for Face Recognition and Clustering" , CVPR 2015,

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed,D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", CVPR 2015

[11] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks", CoRR, abs/1311.2901, 2013

[12] Md. Omar Faruqe, Md. AI Mehedi Hasan, "Face Recognition Using PCA and SVM",2009