A Fusion Face Recognition Approach Based on 7-Layer Deep Learning Neural Network

Jianzheng Liu, Chunlin Fang, and Chao Wu

College of Computer Science and Information Engineering, Tianjin University of Science & Technology, No. 1038, DaGu Road, HeXi District, Tianjin 300222, China

Correspondence should be addressed to Jianzheng Liu; jz.leo@tust.edu.cn

Received 1 December 2015; Accepted 28 March 2016

Copyright © 2016 Jianzheng Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents a method for recognizing human faces with facial expression. In the proposed approach, a motion history image (MHI) is employed to get the features in an expressive face. The face can be seen as a kind of physiological characteristic of a human and the expressions are behavioral characteristics. We fused the 2D images of a face and MHIs which were generated from the same face's image sequences with expression. Then the fusion features were used to feed a 7-layer deep learning neural network. The previous 6 layers of the whole network can be seen as an autoencoder network which can reduce the dimension of the fusion features. The last layer of the network can be seen as a softmax regression; we used it to get the identification decision. Experimental results demonstrated that our proposed method performs favorably against several state-of-the-art methods.

1. Introduction

Face recognition has been one of the hot topics in the biometrics in the past several years. There have been many approaches proposed for this topic. In general, the research of face recognition is focused on verifying or identifying a face from its image.

In 1991, Turk and Pentland present a near-real-time system which tracks a subject’s head and recognizes the person by comparing characteristics of the face to those of known individuals [1]. He projected face images onto a feature space which was named face space. The face space was defined by the “eigenfaces,” which were the eigenvectors of the set faces. This framework provided the ability to learn to recognize new faces in an unsupervised manner. Hence, most researchers’ work was focused on facial feature extraction.

Furthermore, there were some other approaches worthy of attention. In [2] Belhumeur et al. presented an approach based on linear discriminant analysis (LDA). Zhao et al. proposed an approach in [3] which was based on PCA and LDA; they projected the face image to a face subspace via PCA and then LDA is used to obtain a linear classifier. There are also some representative approaches such as locality preserving projections (LPP) which were proposed by He et al. [4] in 2005, as well as marginal fisher analysis (MFA) presented by Yan et al. [5] in 2007.

Some new methods were presented recently. Lu et al. presented an approach in 2013 [6]. Most face recognition methods usually use multiple samples for each subject, but the approach proposed by Lu et al. can work on the condition of only a single sample per person. Yang et al. focused on the speed and scalability of face recognition algorithms [7]. They investigated a new solution based on a classical convex optimization framework, known as Augmented Lagrangian Methods (ALM). Their method provided a viable solution to real-world, time-critical applications. Liao et al. study addresses partial faces recognition [8]. They propose a general partial face recognition approach that does not require face alignment by eye coordinates or any other fiducial points. Wagner et al. proposed a conceptually simple face recognition system [9] that achieved a high degree of robustness and stability to illumination variation, image misalignment, partial occlusion.

These algorithms mentioned above can be summarized into two categories on how to extract facial features: holistic template matching based systems and geometrical local
feature based systems [10]. However, these algorithms utilized only one sort of features in their approach. In [11], a novel multimodal biometrics recognition method was proposed by Yang et al. To overcome the shortcomings of these traditional methods, the proposed method integrated a variety of biological characteristics to identify a subject.

Most classical methods were based on the 2D or 3D images of the face; it could be seen as a kind of physiological characteristic. Human facial expressions are the movements of his facial muscle and can be seen as a kind of behavioral characteristics. All of the characteristics can be used to identify the subject. The research on expression recognition is developing rapidly in recent years, such as the proposed studies in [12–14]. In 2006, Chang et al. proposed a method for 3D face recognition in the presence of varied facial expressions [15]. But in their research, the facial expressions were not used as a kind of characteristics. So in this paper, we propose a face recognition approach based on the fusion of 2D face image and expression. We fuse the physiological and behavioral characteristics in order to improve the performance of our face recognition system. In this paper, we only discuss how to fuse the 2D face images and the expression of "surprise" into a vector to identify the subject. The rest of the paper is organized as follows: in Section 2, we present the relevant theories of our method. Our method is elaborated in Section 3. In Section 4, we report on the experimental setup and results. Section 5 discusses the results of the several experiments, concludes the paper, and outlines directions for future work.

2. Methodology

2.1. Motion Templates. In 1996, motion templates were invented in the MIT Media Lab by Bobick and Davis [16, 17] and were further developed jointly with one of the authors [18, 19]. Initially, motion templates were used to identify the movements of human body by research video or image sequences. For example, Figure 1 shows a schematic representation for a person doing shaking head movement. The algorithm depends on generating silhouettes of the object of interest. There are many ways to get the silhouettes, but it is not the focus of this paper.

Assume that we have an object silhouette. A floating point MHI (motion history image) [20] where new silhouette values are copied with a floating point timestamp is updated as follows:

$$\text{MHI}_\delta(x, y) = \begin{cases} 
\tau & \text{if current silhouette at (x, y)} \\
0 & \text{else if MHI}_\delta < (\tau - \delta).
\end{cases}$$

In the formula, $\tau$ is the current timestamp, and $\delta$ is the maximum time duration constant associated with the template. This method makes the representation independent of system speed or frame rate so that a given gesture will cover the same MHI area at different capture rates [18]. Assume the movement lasted 0.5 seconds. Regardless of the number of pictures in the sequence between the first image and the last one, the area of the MHI was determined by the movement of the head during the time.

2.2. Deep Learning Neural Network. Depth learning neural network has become the most interesting method in the field of machine learning. Essentially, it is the expanding of the traditional neural network approach. The proposed famous training method makes it receive much attention. In 2006, Hinton and Salakhutdinov published a paper in the journal Science [21]. The paper presented a way which can extract features from the raw data automatically. In the paper, Hinton and Salakhutdinov showed a way to reduce a 784-dimensional data (a picture) to a 30-dimensional feature via a 4-layer deep network which was called a DBN (deep belief network). Because the multiplayer neural network was trained by using unsupervised algorithms and can convert high-dimensional data to low-dimensional codes, it is also called “autoencoder.” In general, autoencoder systems were made of more than three layers. The first layer of the network was an input layer, then there were several hidden layers, and the last layer was an output layer.

Each layer of the autoencoder systems is often constituted by RBM (restricted Boltzmann machine). RBMs were invented by Smolensky [22] in 1986. It is a kind of generative stochastic artificial neural network that can learn a probability distribution over its set of inputs. However, the network was growing concern since 2006 until Hinton and Salakhutdinov’s paper was published. The reason they have attracted much attention was because they were used as a single layer to build the multilayer learning systems called deep belief networks, and variants and extensions of RBMs have found applications in a wide range of pattern recognition tasks [23].

An RBM can be seen as an MRF (Markov Random Field) associated with a bipartite undirected graph. The architecture
of an RBM is shown in Figure 2. It is a one-layer network which consists of two kinds of units: \( m \) visible units and \( n \) hidden units. There is a matrix \( W = \{ w_{nm} \} \) that is the weight between the visible units and the hidden units. The visible units correspond to the components of an input vector (e.g., one visible unit for each gray pixel's value of an image). The input data were fed to the network via these units. The hidden units model dependencies between the components of vector (e.g., point, line, or corner in the image) and can be viewed as nonlinear feature detectors [24].

Different with the classical neural network, the deep learning neural networks require more huge samples to train the network. Many deep learning algorithms used unsupervised learning algorithms to train the network. That means the training set was unlabeled samples. It is obvious that capturing unlabeled samples is much easier than making labeled samples. So it is a very important benefit of unsupervised algorithms. The network Hinton and Salakhutdinov used in paper [21] was an example of a deep structure that can be trained in an unsupervised manner. Because of the benefit mentioned above, we used a DBN to solve the face recognition problem in this paper.

3. The Proposed Approach

Figure 3 shows the workflow of our proposed approach. Comparing with other algorithms, our method uses a face image and an MHI (generated from a sequence of images) instead of the single face image as a sample. In order to get the region of interest (ROI), we detect the face in image sequences firstly. After being aligned the ROI images were transformed to gray images and used to generate the MHIs. The MHIs and the features extracted from gray images were resized into vectors and normalized and fused together as features. We therefore used this kind of features to feed a DBN (7-layer deep learning neural network) to recognize faces. The count of input units of the whole network was 20000 and the count of output units was 100.

3.1. Preprocessing. Similar to most of face reorganization algorithms, the first phase of our approach is also locating the face region in images. Plenty of methods can be used to locate faces position. For example, the method proposed by Viola and Jones was a very classical algorithm [25]. How to detect faces in complex background is not the focus of this paper; we used a more smart method via a library proposed by Yu [26], which can detect faces more effectively.

Performance of face recognition systems relies on the accuracy of face alignment and most face alignment methods are based on eye localization algorithm. Hence, we used an eye localization approach proposed by Wang et al. in 2005 [27]. This proposed method had an overall 94.5% eye detection rate and was very close to the manually provided eye positions. Figure 4 is the comparing of the original images and the images after alignment.

3.2. Fusion Features. The aligned face images were resized into a resolution of 100 × 100 pixels. Then we calculated the MHIs on these images. Figure 5 demonstrates an MHI generated from an image sequence which contains an expression "surprise." In the MHI image, it is clear that most pixels' gray value is zero. The pixels which have a nonzero gray value are concentrated in the region of mouth, eyes, and eyebrows. The MHIs contain the facial muscles' movement, such as the region, direction, and speed of a facial organ's movement. Therefore, all the movement can be seen as behavioral characteristics which can be used to identify a subject.

Then we fuse the image features and motion features together. For example, the pixels' gray values in an image can be seen as a kind of features. We selected the first image in a sequence and the MHI calculated from the sequence as raw features. The gray image and the MHI were resized into \( 1 \times 10000 \)-dimensional vectors and normalized into \([-1, 1]\) firstly. Then we connected the two vectors to form a new \( 1 \times 20000 \)-dimensional vector. This vector contains gray image’s information and MHI’s information.

Most face recognition algorithms use single image as a sample, so features extracted from this kind of samples would only contain contour and texture information of a subject. The extracted features can be seen as physiological characteristics. It is difficult to identify subjects by using the facial expression features directly because of the large within-class distance and the small among-class distance. The fusion features contain more information than a single image and have larger among-class distance than expression features. Actually, many kinds of features such as those mentioned in [2–5] can fuse with the expression features and the fused features will improve the performance of the original algorithms.

Finally, all image sequences were transformed into this set of vectors and were fed into the network. The network was made of 7 layers, the previous 6 layers were constituted by RBM, and the last layer was a softmax regression. After training, the trained network would output the recognition results.

3.3. Training Algorithm. Hinton and Salakhutdinov proposed an efficient method in [21] to train a deep belief network. The training process was divided into two stages. The first stage was pretraining phase. In this phase, a stack of RBMs was trained. Each RBM was having only one layer of feature detectors. The outputs of a trained RBM could be seen as features extracted from its input data and were used as the...
In this paper, the proposed method used a 7-layer deep learning neural network, the previous 6 layers were constituted by RBM, and the last layer consisted of a softmax regression. So, we used a different way to train the network. The training process can be divided into three stages. Firstly, we pre-trained the whole network but not the last layer using unsupervised learning algorithm, just like the first phase in Hinton’s method. Because of the benefit of unsupervised learning, we could train the network by feeding a training set consisting of massive unlabeled data to it. After training, the previous 6-layer network could be seen as a features extractor which could extract the most valuable features from the raw data automatically. Then, the last layer was trained in supervised learning with a part of training set and its labels. Lastly, as the way Hinton and Salakhutdinov [21] proposed, the whole network was fine-tuned.

4. Experimental Results

4.1. Training Set and Test Set. Most popular face recognition databases such as the Extended Yale B database do not contain expression image sequences which are not suitable for testing on our method. We therefore established a database in our laboratory. In this database there are 1000 real-color videos in nearly frontal view from 100 subjects, male and female, between 20 and 70 years old. In each section, the subject was asked to make the expression of “surprise” clearly in one time. We captured ten videos from one subject. Each sequence begins with a neutral expression and proceeds to a peak expression, which is similar to the CK+ Expression Database. Each video’s resolution is 640×480 pixels, and the width of a face in frame is in the range of 180 pixels to 250 pixels. All the video clips were captured at 30 frames per second (fps), and each video contained 15 frames. There were four frontal illumination videos, two left side light videos, two right side light videos, and two wearing sunglasses video in the ten videos mentioned above. It should be noted that all samples in our database are using the same pair of sunglasses when they were captured. Figure 6 shows the first frames of the videos in our database.

There are four stages in our experiment. In the first stage, three of the four frontal illumination video segments are used as training samples, and the remaining one was used as test sample. In the second stage, the training set consisted of half of the video clips captured from each subject except the
sunglasses videos. That means for each subject we selected two frontal illumination videos, one left side light video, and one right side light video from his samples. The rest of the videos were used as test samples. In the third stage, the training set consisted of the four frontal illumination videos; the other six videos were used as test samples. We compared PCA features [1], DCT features [28], and fusion features in this stage. In the last stage, we used the same training set in stage 3 to compare our method with the algorithm presented in [8].

4.2. Results. In our previous work, we presented a method for recognizing faces via the same fusion features mentioned in this paper based on principal component analysis (PCA). To show the advantages of the method proposed in this paper, we compared the results here. In the first stage, the method proposed in this paper was yielding 100% recognition accuracy against frontal illumination and the result based on PCA was 97%. Table 1 shows the recognition rate of the second stage. In the table, “front” means frontal illumination, “left” means left side light, “right” means right side light, and “sunglass” means wearing sunglasses. The first row in Table 1 is the results based on PCA; the second row is the results based on the method proposed in this paper which were output by a deep belief network (DBN).

Since our method cannot work on the popular published face database, thus, we did not compare our method with other approaches. However, the rate of our method’s recognition accuracy was a little better than our previous work and some popular methods, such as the reported recognition rate in [29–31].

Table 2 shows the recognition rate of the third stage. The first row in Table 2 is the results based on PCA features, the third row is the results based on DCT features, and the second row and the fourth row are the results based on the fusion features.

Table 3 shows the recognition rate of the fourth stage. The first row in Table 3 is the results based on the method proposed in [8]; the second row is the results based on our method.

4.3. Discussion. Experimental results show that the fusion features proposed in this paper are superior to single features. The MHIs we used to fuse features contain abundant expression characteristics. So there is more information in the fusion features which can be used to identify subjects. In particular, our method has superior performance in the sunglasses tests. The fusion features contain plentiful behavioral characteristics issued by the lower half of human face.

5. Conclusion

In this paper we proposed a novel method for recognizing human faces with facial expression. In order to improve the recognition accuracy rate, we use a kind of fusion features. Normally, the popular methods recognize human face via the image of faces, which can be seen as a kind of physiological characteristic. Facial expressions are the movements of human facial muscles, which also can be used to identify humans faces. It is a kind of behavioral characteristics. Hence, the fusion features contain more information than human face images that can be used for identification and increase the recognition rate especially on illumination variant faces or wearing sunglasses faces.

The approach and experiments presented in this paper are only the preliminary work of our research. In future work we will continuously investigate to try to research the proposed method into two issues. Firstly, we will not only analyze the expression “surprise” emotion in this paper, but also focus on extending our method to all sorts of expression emotions. Actually, we also tested our method with the expression “laugh” in our laboratory. We got a similar recognition accuracy rate which was presented in the paper. But in the condition of wearing scarves, our method did not get a satisfactory result. The reason is that the bottom half of human face contains more expression information than the top half. Secondly, we will study how to combine the popular algorithms into our approach to improve the recognition accuracy rate.

Competing Interests

The authors declare that they have no competing interests.
Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant no. 61502338, the 2015 key projects of Tianjin Science and Technology Support Program no. 15ZCZDGX00200, and the Open Fund of Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis no. GDUPTKLAB201334.

References


