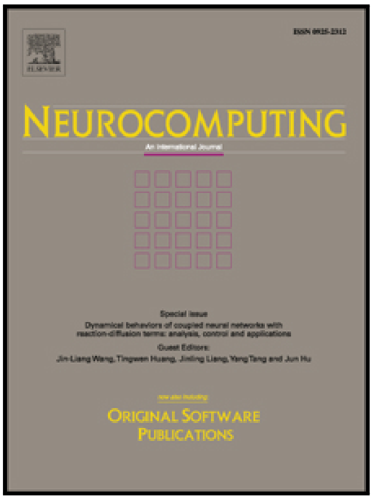
Communicated by Prof. Liu Guangcan



Face Recognition based on Recurrent Regression Neural Network

Yang Li, Wenming Zheng, Zhen Cui, Tong Zhang

PII: S0925-2312(18)30172-3
DOI: 10.1016/j.neucom.2018.02.037
Reference: NEUCOM 19327

To appear in: *Neurocomputing*

Received date: 6 July 2017
Revised date: 14 October 2017
Accepted date: 6 February 2018

# Face Recognition based on Recurrent Regression Neural Network

Yang Li[a,b], Wenming Zheng[a,*], Zhen Cui[c,*], Tong Zhang[a,b]

[a] Key Laboratory of Child Development and Learning Science of Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing, Jiangsu 210096, China
[b] Department of Information Science and Engineering, Southeast University, Nanjing, Jiangsu 210096, China
[c] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210096, China

## Abstract

To address the sequential changes of images including poses, in this paper we propose a recurrent regression neural network (RRNN) framework to unify two classic tasks of cross-pose face recognition on still images and videos. To imitate the changes of images, we explicitly construct the potential dependencies of sequential images so as to regularizing the final learning model. By performing progressive transforms for sequentially adjacent images, RRNN can adaptively memorize and forget the information that benefits for the final classification. For face recognition of still images, given any one image with any one pose, we recurrently predict the images with its sequential poses to expect to capture some useful information of other poses. For video-based face recognition, the recurrent regression takes one entire sequence rather than one image as its input. We verify RRNN in still face image dataset MultiPIE and face video dataset YouTube Celebrities (YTC). The comprehensive experimental results demonstrate the effectiveness of the proposed RRNN method.

*Keywords:* Recurrent regression neural network (RRNN), face recognition, deep learning

[*] Corresponding author
*Email addresses:* wenming_zheng@seu.edu.cn (Wenming Zheng), zhen.cui@njust.edu.cn (Zhen Cui)

## 1. Introduction

Face recognition is a classic topic in past decades and now still attracts much attention in the field of computer vision and pattern recognition. Face recognition has a great potential in multimedia applications, e.g. video surveillance, personal identification, digital entertainment and so on [1, 2, 3, 4, 5]. With the rapid development of electric equipment techniques, more and more face images can be easily captured in the wild, especially videos from cameras of surveillance or cell phones. Therefore, video or image set based face recognition becomes more important in most of real-world applications and also becomes a popular topic in face analysis more recently. As face images captured from the unconstrained conditions are usually with complex appearance variations in poses, expressions, illuminations, *etc.*, the existing face recognition algorithms still suffer from a severe challenge in fulfilling real applications to large-scale data scenes, although the current deep learning techniques have made a great progress on the unconstrained small face dataset, *e.g.*, the recent success of deep learning methods on Labeled Faces in the Wild (LFW) [6].

In the task of face recognition, however, we cannot bypass this question of pose variations, which has been extensively studied and explored in past decades, and has not been well-solved yet. The involved methods may be divided into 3D [7, 8, 9] and 2D methods [10, 11, 12, 13, 14, 15, 16]. Since pose variations are basically caused by 3D rigid motions of face, 3D methods are more intuitive for pose generation. But 3D methods usually need some 3D data or recovery of 3D model from 2D data which is not a trivial thing. Moreover, the inverse transform from 3D model to 2D space is sensitive to facial appearance variations. In contrast to 3D model, due to decreasing one degree of freedom, 2D methods usually attempt to learn some transforms across poses, including linear models [17] or non-linear models [10, 18]. Because of its simplicity, 2D model has been widely used to deal with cross-pose face recognition with a comparable performance with 3D model. However, in many real scenes of face image sets, *e.g.*, face video sequences, the changes of poses may be regarded as a nearlycontinuous stream of motions, while the existing methods usually neglect or do not make full use of this prior. Moreover, the pose variation is not the only factor between different images even for the same subject, which involves other complex factors.

In this paper, we propose a recurrent regression neural network (RRNN) framework to explicitly construct the potential dependencies of sequential images and unify two classic tasks of face recognition, i.e., face recognition based on still images and videos, respectively. For face recognition of still images, given any one image with any one pose, we recurrently predict the images with its sequential poses to capture some useful information of other poses, under the supervision of known pose sequences. For video-based face recognition, we deal with the recognition problem from entire sequence rather than one image used in still images based face recognition. In detail, by repetitively regularizing the relationship of adjacent frames, we can obtain more robust representation of face video sequences under the supervised case. RRNN can adaptively memorize and forget the information that benefits for the final classification through continuously transferring information from sequentially adjacent images.

The major contributions of this paper can be summarized as follows:

- We construct potential dependencies of sequential images. Benefit from that, the proposed recurrent regression neural network (RRNN) captures the different poses information adaptively for face recognition;
- By constructing a virtual sequence, RRNN makes sense with two different face recognition tasks, i.e., still image and video-based face recognition.

The rest of the paper is organized as follows. We introduce preliminary works in the Section 2 including some fundamental knowledge about face recognition across poses, video-based face recognition and recurrent neural networks (RNNs). In Section 3, we present the proposed Recurrent Regression Neural Network (RRNN) model and its applications to classify still images and video sequences. Section 4 depicts the experiments and discussion. At last, we make a conclusion of this paper.

## 2. Preliminaries

### 2.1. Face Recognition across Poses

With the development of 3D camera technology, several researches try to solve face recognition problem with 3D face images. 3D face images can solve the problem that the distance between two certain parts of face varying in different poses, and 3D face images contain more information about the face such as the depth of the facial features. The current 3D technologies include 3D images processing (captured by 3D camera) and 3D recovery (transformed from 2D to 3D images (2D→3D)). For example, Drira et al. [8] proposed a novel geometric framework for analyzing 3D faces. It represents facial surfaces by radial curves emanating from the nose tips and uses elastic shape analysis of these curves to develop a Riemannian framework for analyzing shapes of full facial surfaces. Asthana et al. [9] proposed a 3D pose normalization method that is completely automatic and leverages the accurate 2D facial feature points found by the system. Li et al. [19] proposed a novel method, named Morphable Displacement Field (MDF), using a virtual view to match the pose image.

3D technology has been confirmed to own a good performance in face recognition [8, 9]. However, 3D data is hard to get in some unconstrained scenes, and the 2D to 3D algorithms are still needed to be explored.

Meanwhile, there have been existed a lot of algorithms for traditional 2D face image recognition. For pose variation face recognition, Sharma et al. [20] linearly mapped images in different modalities to a common linear subspace in which they are highly correlated. And then they presented a general multi-view feature extraction by learning a common discriminative subspace, in which pose variation is minimized [21]. Kan et al. [10] proposed a stacked progressive auto-encoders network, which changes the larger poses to frontal pose.

### 2.2. Video-based Face Recognition

Video-based face recognition is generally studied for three scenes, namely Still-Video, Video-Still and Video-Video [22]. Still-Video face recognition searches the still image in a video. It is always used to find a man in a video when given only a face image of him. On the contrary, Video-Still face recognition matches a video of a man in a lot of images. Video-Video inquires the clip of a man's video given in a lot of videos. Video-based face recognition is different from face recognition of still image. Under normal circumstances, the faces captured by a camera are always affected by the environment around seriously and sometimes have low quality, *e.g.*, large angle pose, blur, low resolution and complex illumination. Consequently, video-based face recognition is more challenging than still image face recognition. Especially, we noticed that face poses are different in each frame of a video because head in videos usually swings around.

Benefiting from the good performance of 2D face recognition technology, video-based face recognition causes several researchers' attention. Hadid et al. [23] proposed a novel approach based on manifold learning to solve the problem of video-based face recognition in which both training and testing sets are video sequences. Chen et al. [24] introduced the concept of video-dictionaries for face recognition, which generalizes the work in sparse representation and dictionaries for faces in still images.

### *2.3. Recurrent Neural Networks (RNNs)*

Recurrent Neural Networks (RNNs) are popular models, which have shown great performance in many tasks [25, 26]. The idea behind RNNs is to make use of the sequential information by mapping input sequence to a sequence of hidden states, which can learn the complex dynamics of sequence. The recurrence equations from the hidden states to outputs is as follows:

$$\mathbf{h}_t = \sigma(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h), \tag{1}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{ho}\mathbf{h}_t + \mathbf{b}_o), \tag{2}$$

where $\mathbf{x}_t$ is the $t$-th input of the sequence $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T\}$ with length $T$. The corresponding hidden states and outputs are $\{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_t, \cdots, \mathbf{h}_T\}$ ($\mathbf{h}_0 = \mathbf{0}$) and $\{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_t, \cdots, \mathbf{o}_T\}$ respectively. $\sigma$ is an element-wise nonlinear activation function. $\mathbf{W}_{xh}, \mathbf{W}_{hh}$ and $\mathbf{W}_{ho}$ are transform matrices, and $\mathbf{b}_h$

and $\mathbf{b}_o$ is the biases [27]. By using these recurrent operations, RNNs capture the sequential and time dependencies via cycles in the network of nodes [28].

## 3. Recurrent Regression Neural Network

In this section, we first provide an overview introduction on the proposed recurrent regression neural network (RRNN) framework, then two cases of face recognition based on cross-pose and video are further modeled.

### *3.1. The Model of RRNN*

The overall framework of RRNN is shown in Fig. 1. To make full use of various appearance models for still image or video-based face recognition, we explicitly build a recurrent regression model to transform the current input into other appearance spaces, in which we seek for some effective components to compensate mismatching appearance variations of face images. Given an input $\mathbf{x}_i$, we encode it into a latent state $\mathbf{S}_i$ and then decode it to one virtual output $\widetilde{\mathbf{x}}_i$, which may come from the other space spanned by some other appearance characteristics we expect. The encoder-decoder models a dynamic changing process between the input and the expected output, and may be further stacked layer by layer to represent a sequence, *i.e.*, a process of recurrent encoding-decoding. In the task of face recognition, in order to enhance the model discriminability, the identification of subjects can be combined into this model as a joint learning. Besides, to reduce error-drifting of all decodings, we impose a total error constraint on the sum of all outputs to explicitly smooth the entire output sequence. Concretely, we formulate recurrent regression on a sequence of appearance variations from three aspects:

(1) Recurrent encoding-decoding. Let $\{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{t-1}, \mathbf{x}_t, \cdots\}$ denote the input sequence states, then the corresponding hidden states $\{\mathbf{S}_0, \cdots, \mathbf{S}_t, \cdots\}$ after encoding can be written as

$$\mathbf{S}_t = \sigma(\mathbf{U}g(\mathbf{x}_t) + \mathbf{W}\mathbf{S}_{t-1} + \mathbf{b}_1), \tag{3}$$

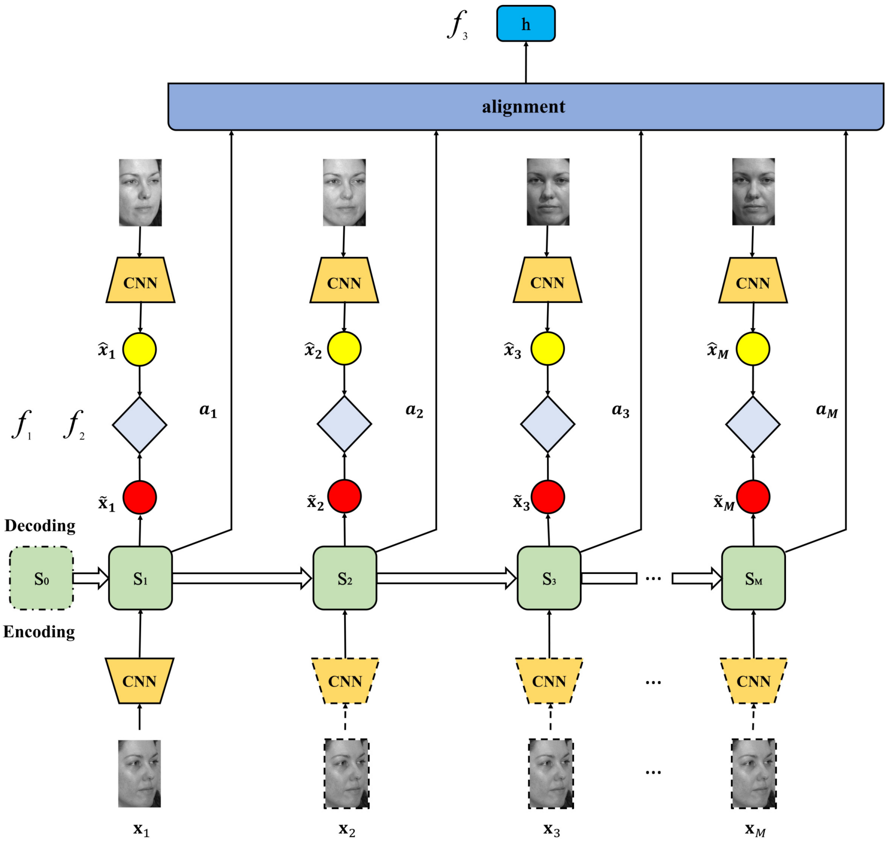


Figure 1: An illustration of our idea. Given one pose-specified face image $\mathbf{x}_1$, we use a regularized recurrent neural network (RNN) to progressively regress the pose-stream images $(\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_M)$ so that pose mismatching in face recognition can be reduced. In the general case of still image (as shown in this figure), the only one input image is simply replicated as a virtual sequence for the input of RNN. The sequential hidden states of different poses are adaptively weighted to form the final face representation. To increase the feature discriminability, the CNN network is used to extract more abstract high-level features. Here the rhombuses are the operations between $\widehat{\mathbf{x}}_\mathbf{t}$ and $\widetilde{\mathbf{x}}_\mathbf{t}$, i.e., Eq. (5).

where $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are linear transform matrices, $\mathbf{b}_1$ is the bias term, and $\sigma$ is a nonlinear transform activation function, *e.g.*, the Hyperbolic Tangent (tanh) function used in this paper. $g$ is the nonlinear transform by CNN while $g(\mathbf{x}_t)$ is the representation of $\mathbf{x}_t$. Note that, the encoding also depends on the previous hidden state $\mathbf{S}_{t-1}$ partly besides the current input because previous historic information may bring external beneficial information for the next representation. Further, we decode each hidden state $\mathbf{S}_t$ into its specified output $\widetilde{\mathbf{x}}_t$ we expected, *i.e.*,

$$\widetilde{\mathbf{x}}_t = \sigma(\mathbf{V}\mathbf{S}_t + \mathbf{b}_2), \qquad t = 1, 2, \cdots, \tag{4}$$

where $\mathbf{V}$ is the decoding matrix and $\mathbf{b}_2$ is the bias. Consequently, this objective function is to minimizing all reconstruction errors, *i.e.*,

$$f_1 = \sum_{t=1,2,\cdots} \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|_F^2, \tag{5}$$

where $\widehat{\mathbf{x}}_t$ is the ground-truth of the next state in the sequence.

(2) Sequence reconstruction. Let $\{\mathbf{x}_1^o, \mathbf{x}_2^o, \cdots, \mathbf{x}_t^o, \cdots\}$ denotes the original sequence with all poses. To further characterize the globality of the decoding on a sequence, we force some statistic properties of all outputs to close to be an expected state, *i.e.*, minimizing the following objective function,

$$f_2 = \sum_{t=1,2,\cdots} \|f(\widetilde{\mathbf{x}}_t) - f(g(\mathbf{x}_t^o))\|_F^2, \tag{6}$$

where $f$ is the statistic function on a sequence, such as first-order statistics, i.e., mean value, etc.

The major differences between $f_1$ and $f_2$ can be summarized as follows:

- From the functions' point of view, $f_1$ is used to force the decoded unit to the next pose image and finally it changes the decoded unit to the frontal pose, which is easier to be classified in face recognition. But $f_2$ forces the statistics of input to an expected state which comes from the original sequence with all poses.

- From the properties' point of view, $f_1$ aims at pose transformation separately. But $f_2$ is designed to utilize the global information of all poses, and collaborate all reconstruction units to reduce error-propagation, which can further improve the performance according to our observation from the following experiments.

(3) Discriminative prediction. Like most supervised models, we may add a supervision term into the network so as to enhance the model discriminability. Concretely, we use softmax function on the transformed hidden states, *i.e.*,

$$\mathbf{O}_t = \mathbf{G}\mathbf{S}_t + \mathbf{b}_3 = [O_{t,1}, O_{t,2}, \cdots, O_{t,i}, \cdots], \tag{7}$$

$$P_t(y\!=\!i|\mathbf{S}_t, \mathbf{G}, \mathbf{b}_3) \!=\! \frac{\exp(O_{t,i})}{\sum_{c=1,2,\cdots,i,\cdots} \exp(O_{t,c})}, \tag{8}$$

$$f_3 = -\sum_t \log(P_t(y = i|\mathbf{S}_t, \mathbf{G}, \mathbf{b}_3)), \tag{9}$$

where the variables $\mathbf{G}$ and $\mathbf{b}_3$ are respectively the transform matrix and the bias. $i$ is the $i$-th class. Note that here the supervision information is directly imposed on the hidden states rather than the decoding output $\widetilde{\mathbf{x}}_t$. The reasons are two folds: i) the reconstruction in each decoding unit is not perfect, where the errors might reduce the discriminative capability especially when accumulatively propagated along the sequential network; ii) it can implicitly transit some identification information to the reconstruction stage and thus reduce the direct influence on decoding targets due to the large semantic gap between reconstruction targets and labels.

Furthermore, to further characterize the globality of hidden states, we adaptively weight each hidden state by introducing a subnetwork called hidden state alignment. This subnetwork globally balance all hidden states to decide the label of sequence. The concrete process can be written as

$$\mathbf{h} = \sum_t \mathbf{a}_t \mathbf{S}_t, \tag{10}$$

$$\mathbf{O} = \mathbf{G}\mathbf{h} + \mathbf{b}_3 = [O_1, O_2, \cdots, O_i, \cdots], \tag{11}$$

$$P(y{=}i|\mathbf{h}, \mathbf{G}, \mathbf{b}_3) {=} \frac{\exp(O_i)}{\sum_{c=1,2,\cdots,i,\cdots} \exp(O_c)}, \tag{12}$$

$$f_3 = -\log(P(y = i|\mathbf{h}, \mathbf{G}, \mathbf{b}_3)), \tag{13}$$

where $\mathbf{a}_t$ is a weight vector to alignment the hidden state $\mathbf{S}_t$.

Finally, the overall objective function can be defined as

$$\min \quad f_1 + \alpha f_2 + \beta f_3, \tag{14}$$

where $\alpha$ and $\beta$ are the balance parameters corresponding to the sequence reconstruction term and the supervision term. To make it easier to be understand, we mark the three loss function in Fig. 1 at their calculation position.

### *3.2. RRNN for Still Images across Poses*

For still images with different poses, the poses can be sorted in a sequence according to the continuity of pose changing. In the training stage, the image sequence along pose changes may be easily captured from cameras, and thus can be used in the proposed recurrent regression network. However, in the classic task of cross-pose face recognition based on still images, only one image is usually provided for testing. So we have to flatter this model by converting one image into a virtual sequence.

Given any one image $\mathbf{x}_1^o$ with one pose, we augment it into a sequence stream by using repeatedly copy, *i.e.*, $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t, \cdots\} = \{\mathbf{x}_1^o, \mathbf{x}_1^o, ..., \mathbf{x}_1^o, \cdots\}$, which is pretended to be the input sequence. For the decoding outputs, we expect to predict those images of other poses. In order to utilize gradual changes of poses, we construct the decoding output sequence as the adjacent pose stream $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_t, \cdots\} = \{g(\mathbf{x}_2^o), g(\mathbf{x}_3^o), ..., g(\mathbf{x}_{t+1}^o), \cdots\}$, *i.e.*, the next adjacent pose is its decoding output, where $g$ is the CNN operations. In this way, recurrent encoding-decoding can realize the function that transforms the input pose to the target pose we expect. For the sequence reconstruction term, we use the mean values of predicted state as first-order statistics, and then make it close to the mean of pose streams. This term could collaborate all reconstruction poses to reduce error-propagation and also take advantage of the global information.

### *3.3. RRNN for Video Sequences*

Different from the case above in Section 3.2, for video-based face recognition, the input sequence is explicitly known in the testing stage. For RRNN, thus the input sequence consists of all frames of a video sequence, *i.e.*, $\{\mathbf{x}_1^o, \mathbf{x}_2^o, ..., \mathbf{x}_t^o, \cdots\}$, where $\mathbf{x}_t^o$ is the $t$-th frame of the sequence. Instead of the use of next frames, we use the mean value of all frames as the decoding outputs we expect, *i.e.*, $\{g(\overline{\mathbf{x}}), g(\overline{\mathbf{x}}), \cdots, g(\overline{\mathbf{x}}), \cdots\}$, where $\overline{\mathbf{x}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t^o$. The main reason is that our aim is to classify each sequence rather than predict next frames. If we use next frames as the decoding outputs, we could capture more motion information in the encoding-decoding process, which does not refine the subject information yet. Under this constraint of mean prediction, the sequence reconstruction term will play an unimportant role in the final performance also as observed from our experiments, due to the nearly-common optimization target.

## 4. Experiments

### *4.1. Experimental Set-up*

In this section, we evaluate our proposed RRNN on two widely used face datasets, one is the cross-pose face dataset MultiPIE [29], and the other is the video dataset YouTube celebreties (YTC) [30]. As Convolutional Neural Network (CNN) can extract more robust features according to the recent researches, so in this experiment we employ CNN features to represent the images to feed into RRNN as inputs. Concretely, we directly employ the released training model of VGGFACE [31] network to extract face features, where images are first up-scaled into 256*256 and the output of 2622 dimension on the layer ‘fc8’ are used as the feature of each image. VGGFACE comprises 11 blocks and each block contains a linear operator followed by one or more non-linearities such as ReLU and max pooling. There are totally 16 convolution layers in VGGFACE. Of course, the CNN model can be concatenated with our RRNN for an end-to-end neural network. Considering small scale training samples, we only use it to extract features to verify our idea. Without fine tuning on the network parameters, we simply set the number of hidden units to 5000 as default value in the following experiments. In the training process of our experiments, we stop training at the time of the model scans the training data 100 times without repetition. RRNN is implemented with the popular Theano.

### *4.2. Experiment of Face Recognition across Poses on MultiPIE Database*

MultiPIE dataset contains 337 people with face images of different poses, illumination and expressions. Each person has 7 poses from $45^\circ$ to $-45^\circ$ with $15^\circ$ interval, where $0^\circ$ means the frontal pose. Following the same experiment configuration of [10], we choose the first 200 subjects (subject ID from 1 to 200) as the training set, totally 4207 face images. The rest 137 subjects are used as the testing set, totally 1879 face images. Inside the testing set, we take one frontal pose face image from each subject, totally 137 frontal pose face images as the gallery set. The rest 1742 face images are used as probe set.

In MultiPIE dataset, each individual ID and poses are known in the training set. Thus, in the training process, we can train our RRNN model to make it the ability to force the input facial image to the frontal pose facial image by using objective function $f_1$ in Eq. (5), which is better to be recognized in face recognition task. But in the testing process, we don't know the pose and ID of all individuals in the test set. And in face recognition for still image, only one image with an unknown pose is given to be recognized. Hence, we augment this image into a sequence by repeating copy, *e.g.*, $\{\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_t; \cdots\} = \{45^\circ; 45^\circ; \cdots; 45^\circ; \cdots\}$, so that we construct a virtual sequence as the input sequence for still image face recognition and thus make sure the input of the model has the same formulation in training and testing process.

As face images with $0^\circ$ pose are easier to be recognized according to human cognition, we convert each pose to the frontal pose by using the gradual changing strategy. For example, given an image of $-45°$ pose, we expect the decoding sequence to be faces with $\{-30°, -15°, 0°\}$ poses. However, images of different poses will have a regression sequence with different lengths. To handle this problem, we pad the frontal pose into those short sequences so as to generate the encoding sequences with equal length. In order to identify the end of front pose in the testing stage, we externally extend the sequence by adding the frontal pose at the end of each sequence, where we expect the front pose will not be changed in the looped regression model as the terminate state.

As the subject labels contain weak discriminative information in face recognition, thus, although training set and the gallery set do not share the same label information, we also use the discriminative term in RRNN, which in fact directly verifies our idea of recurrent regression itself. Given a testing sample, we take the values of each hidden state as the regressed features to construct a similar matrix, and employ idea of Nearest Neighbor (NN) algorithm to classify it. For RRNN, we set the balance parameter $\alpha = 1$ and $\beta = 10$ in Eq. (14). Table 1 shows the results.

In order to test the performance of our RRNN, we compare it with the state-of-the-art methods, consisting of 3D and 2D technologies. For 3D technologies, we compare the two methods of Asthana11 [9] and MDF [19]. For 2D technologies, we compare those classic models including PLS [20], CCA [32], GMA [21], DAE [33], SPAE [10], LDA-SID and DFD-SID [34], where LDA-SID achieves the current best performance on this protocol. To verify the effectiveness of the regression model in principle, we also conduct the experiment VGG+KNN. The comparison results are reported in Table 1, we can have two main observations from it:

1) From the eighth line of this table, VGG [31]+KNN achieves a competitively performance compared with the existing methods, even the recently proposed deep learning method SPAE. It again indicates that CNN can benefit face recognition more than those raw/hand-crafted features.

Table 1: The classification results on MultiPIE dataset.

| Methods | Probe Pose | | | | | | |
|---|---|---|---|---|---|---|---|
| | $-45°$ | $-30°$ | $-15°$ | $+15°$ | $+30°$ | $+45°$ | Average |
| Asthana11 [9] | 74.1% | 91.0% | 95.7% | 95.7% | 89.5% | 74.8% | 86.8% |
| MDF [19] | 78.7% | 94.0% | 99.0% | 98.7% | 92.2% | 81.8% | 90.7% |
| PLS [20] | 51.1% | 76.9% | 88.3 % | 88.3% | 78.5 % | 56.5% | 73.3% |
| CCA [32] | 53.3% | 74.2% | 90.0% | 90.0% | 85.5% | 48.2% | 73.5% |
| GMA [21] | 75.0% | 74.5% | 82.7% | 92.6% | 87.5% | 65.2% | 79.6% |
| DAE [33] | 69.9% | 81.2% | 91.0% | 91.9% | 86.5% | 74.3% | 82.5% |
| SPAE [10] | 84.9% | 92.6% | 96.3% | 95.7% | 94.3% | 84.4% | 91.4% |
| VGG+KNN | 83.0% | 94.9% | **98.6%** | 97.9% | 94.6% | 85.8% | 92.5% |
| LDA-SID [34] | **92.3%** | 96.0% | 98.0% | 96.7% | 94.7% | 91.0% | 94.8% |
| DFD-SID [34] | 91.3% | 95.3% | 97.7% | 96.3% | 94.3% | 90.0% | 94.2% |
| **RRNN** | 91.1% | **97.7%** | 98.3% | **98.3%** | **97.6%** | **91.3%** | **95.6%** |

2) Although CNN features are robust enough, RRNN can further improve the performance by using the prior of pose changing. Our RRNN performs best in poses of $-30°$, $15°$, $30°$ and $45°$. Compared to LDA-SID, of which results are recently released, the average improvement is about 0.8 percent.

### *4.3. Experiment of Video-Video Face Recognition on YTC Database*

YouTube celebreties (YTC) [30] dataset contains 1910 face videos of 47 people. These videos are with large variation of pose, illumination and expression. As compression ratio of most videos is very high, the quality of faces in video are usually very poor, especially including some factors of blur, low-resolution, fast motion, *etc*. Furthermore, the number of video frames ranges from 7 to 400.

As described in [22], we detect the faces in YTC videos and align them into $20 \times 20$. Following the protocol in [22] strictly, we randomly choose a video of each session of each subject for training, and choose 2 videos from the rest videos for each session for testing. There are total 3 sessions, thus 3 samples of each subject are used for training and 6 samples for testing. Ten trials are randomly conducted so as to cover all samples. The average accuracy of ten trials is used as the final result. In this training, to reduce the computation cost of each sequence and increase the training sequences, we cut each video to several clips of 10 frames. And for a testing sequence, we vote the label of all clips of a sequence as the final label. Due to the shared labels of training set and testing set, we use the discriminative model (*i.e.*, logistic regression) to predict the classification score.

Different from the case of still face image dataset MultiPIE, poses are unknown in YTC dataset. But in face recognition task, our aim is to classify the sequence rather than to predict the next frame. Because the above two reasons, we treat the original video sequence as the input sequence, which can also utilize the information of all the poses, and use the mean value of all frames repeated as the decoding sequence in objective function $f_1$ in Eq. (5), i.e., $\{\widehat{\mathbf{x}}_1; \widehat{\mathbf{x}}_2; \cdots; \widehat{\mathbf{x}}_t; \cdots\} = \{\mathbf{x}_1^o, \mathbf{x}_2^o, ..., \mathbf{x}_t^o, \cdots\}$, where $\mathbf{x}_t^o$ is the $t$-th frame of the data sequence, and discard objective function $f_2$ due to the nearly-common optimization target. By doing the above, we don't need to modify the model and perform face pose alignment additionally but the experimental results show good performance of our model.

Here we set $\beta = 1$ and compare RRNN with several state-of-the-art algorithms, including MSM [35], DCC [36], MMD [37], MDA [38], AHISD [39], CHISD [39], SANP [40], CDL [41], DFRV [24], LMKML [42], SSDML [43], SFDL [44] and MDML [22]. Their mean accuracies are reported in Table 2. These methods fall into the category of subspace based or metric based methods. It is apparent RRNN gets the best performance compared with all the other algorithms. Furthermore, the improvement is up to 6.1%. This huge improvement indicates RRNN can well model video sequence.

Table 2: Average classification result and standard deviation on YTC dataset.

| Methods | MSM[35] | DCC[36] | MMD[37] | MDA[38] |
|---|---|---|---|---|
| YTC | 61.7±4.3 | 65.8±4.5 | 67.7±3.8 | 68.1±4.3 |
| Year | 1989 | 2006 | 2008 | 2009 |
| Methods | AHISD[39] | CHISD[39] | SANP[40] | CDL[41] |
| YTC | 66.5±4.5 | 67.4±4.7 | 68.3±5.2 | 69.7±4.5 |
| Year | 2010 | 2010 | 2011 | 2012 |
| Methods | DFRV[24] | LMKML[42] | SSDML[43] | SFDL[44] |
| YTC | 74.5±4.5 | 75.2±3.9 | 74.3±4.5 | 75.7±3.4 |
| Year | 2012 | 2013 | 2013 | 2014 |
| Methods | MDML[22] | VGG+DARG-KLD [45] | VGG+DARG-MD+LED[45] | RRNN |
| YTC | 78.5±2.8 | 79.9 ±1.7 | 83.4±1.4 | **84.6**±2.1 |
| Year | 2015 | 2015 | 2015 | |

For Multi-PIE, we provide a standard baseline (VGG features) to verify the effectiveness of our idea. Similarly, we conduct an extra experiment on YTC based on the VGG features. We choose the current best conventional method on YTC, called DARG-KLD and DARG-MD+LED [45], and use their released code. The same VGG features are fed into DARG-KLD and DARG-MD+LED. Its performance is 79.9% and 83.4% (vs. ours 84.6%). Moreover, we also conduct t-test statistical analysis with the significant level $\alpha = 0.05$ for the experimental result to see whether RRNN has an improvement of recognition rate compare with the baseline methods. We assume that the mean classification result of RRNN is same with DARG-KLD and DARG-MD+LED, and the rejection region is t≥ $t_{0.05}(10 + 10 - 2) = t_{0.05}(18) = 1.734$. It means that RRNN can achieve a better result than the corresponding compared algorithms when the t-value falls into this region. Table 4.3 shows the t-test statistical analysis results. From Table 4.3, we can see RRNN is significantly better than the baseline methods. Furthermore, our RRNN is more simplified but effective.

Table 3: The statistics analysis of RRNN and the baseline methods.

| Method | t-value |
|---|---|
| RRNN vs. VGG+DARG-KLD | 4.336* |
| RRNN vs. VGG+DARG-MD+LED | 1.802* |

* RRNN is significantly better than the compared algorithm

Table 4: The performance of three objective functions on MultiPIE dataset.

| Objective function | $f_1$ | $f_2$ | $f_3$ | $f_1 + f_2$ | $f_1 + f_3$ | $f_2 + f_3$ | $f_1 + f_2 + 10f_3$ |
|---|---|---|---|---|---|---|---|
| Accuracy | 94.7% | 94.6% | 91.5% | 95.2% | 95.0% | 94.9% | **95.6%** |

### *4.4. Discussion*

#### *4.4.1. The effectiveness of terms in the objective function*

As described in the objective function Eq. (14), there are two related parameters $\alpha$ and $\beta$, which respectively constrain the sequence reconstruction and label prediction.

As Table. 4 shows, according to the above analysis, in the task of face recognition across poses, with recurrent encoding-decoding $f_1$, the recognition accuracy is up to 94.7%. By adding sequence reconstruction $f_2$, the performance is further improved up to 95.2% when $\alpha = 1$. Furthermore, by adding discriminative prediction $f_3$, the performance is further improved up to 95.6% when $\beta$ equals 10. Meanwhile, we treat the still image recognition settings as video-based face recognition, i.e., only use $f_1$ and $f_3$. The performance is improved about 0.3% compared with only use $f_1$. This indicate the ID information in training set can brings some discrimination in still image face recognition. But the result of using $f_3$ as the objective function only gets a worse performance compared with that of using $f_1$ and $f_2$ separately only and the gap is about 3%. This shows image features, i.e., the recurrent encoding-decoding $f_1$ and sequence reconstruction item $f_2$, are more useful than the ID information, i.e., discriminative prediction $f_3$ in still image face recognition task.

In the task of video-based face recognition, we discard the sequence reconstruction term $f_2$ due to the common target with the encoding-decoding process as analyzed above. By adding the label loss term $f_3$, we can promote the performance about 1 percent when $\beta = 1$.

For $\alpha$ and $\beta$, we only tune them in the range $\{1, 10\}$ without finer tuning. Even though, we find they benefit the final classification performance by introducing the learning of appearance variations.

#### *4.4.2. Cross-pose analysis*

Although in the above experiments on MultiPIE the frontal pose is specified as the terminate state of the proposed recurrent regression model, we can directly obtain all cross-pose results based on this frontal pose model by selecting a pose as the gallery set and the rest poses as the probe set. Table 5 shows the results of different poses as gallery sets for face recognition across poses on MultiPIE. It is interesting to observe that the recognition result doesn't reach the best when images of frontal pose ($0°$) are used as the gallery set, which seems not to match with our intuition. The reasons should come from two aspects: 1) Frontal faces in the gallery set are currently decoded along the time stream, and reconstruction errors are propagated with the evolution of states, which leads to a derivation from the ground-truth frontal faces for the decoding states. 2) According to the symmetry of faces, non-frontal faces can induce frontal faces to some extends as non-frontal faces contain more contour information than frontal faces.

#### *4.4.3. The effectiveness of hidden states alignment*

In Eq.(14), there are two computation methods, i.e., Eq.(9) and Eq.(13), for $f_3$. To evaluate the effectiveness that the hidden states alignment can find out a deep relationship among different hidden states, we get two results of 84.3% and 84.6% on YTC dataset using Eq.(9) and Eq.(13) respectively. It is a slight promotion with hidden states alignment due to that the hidden states in RRNN have been trained fine to get a better feature of each input. But it is necessary

Table 5: Cross-pose results of our proposed RRNN on MultiPIE.

| | | Probe Pose | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $-45°$ | $-30°$ | $-15°$ | $0°$ | $+15°$ | $+30°$ | $+45°$ | Average |
| Gallery Pose | $-45°$ | - | 0.9825 | 0.9790 | 0.9694 | 0.9651 | 0.9755 | 0.9818 | 0.9756 |
| | $-30°$ | 1.0000 | - | 1.0000 | 0.9926 | 0.9966 | 0.9932 | 0.9746 | 0.9929 |
| | $-15°$ | 0.9889 | 1.0000 | - | 1.0000 | 1.0000 | 1.0000 | 0.9746 | 0.9939 |
| | $0°$ | 0.9118 | 0.9732 | 0.9833 | - | 0.9833 | 0.9764 | 0.9130 | 0.9568 |
| | $+15°$ | 0.9705 | 1.0000 | 1.0000 | 1.0000 | - | 1.0000 | 0.9782 | 0.9915 |
| | $+30°$ | 0.9705 | 0.9966 | 1.0000 | 1.0000 | 1.0000 | - | 0.9963 | 0.9939 |
| | $+45°$ | 0.9738 | 0.9752 | 0.9717 | 0.9615 | 0.9717 | 0.9893 | - | 0.9739 |

to find out a deep relationship among all the hidden states.

### *4.4.4. Parameter analysis*

Here, we analyze the performance of different parameters $\alpha$ and $\beta$ in the experiment on MultiPIE dataset. For deep learning methods, the choice of parameters is an unsolved problem and is not a topic in this paper. Thus, in our experiment, first we set $\alpha = 1$ simply and induce Fig. 2 that shows the performance of different $\beta$ when $\alpha = 1$. In Fig. 2, we can see that the best accuracy is 95.6% when $\beta = 10$. Then, the performance of different $\alpha$ when $\beta = 10$ is shown in Fig. 3. From Fig. 2 and Fig. 3, we can see that the best performance appears when $\alpha = 1$ and $\beta = 10$.

### *4.4.5. Convergence analysis*

The convergence of deep learning methods is still unsolved and beyond our topic. Here we simply depict the curves of objective function value and classification rate to show the convergence of RRNN. Fig. 4 shows the objective function loss (Eq.(14)) versus different number of iterations on MultiPIE dataset. It is easy to see our RRNN fast converges a local minimum after a few iterations. Fig. 5 shows the classification rate versus different number of iterations on MultiPIE dataset. We can see that our RRNN converges in 30 iterations and the

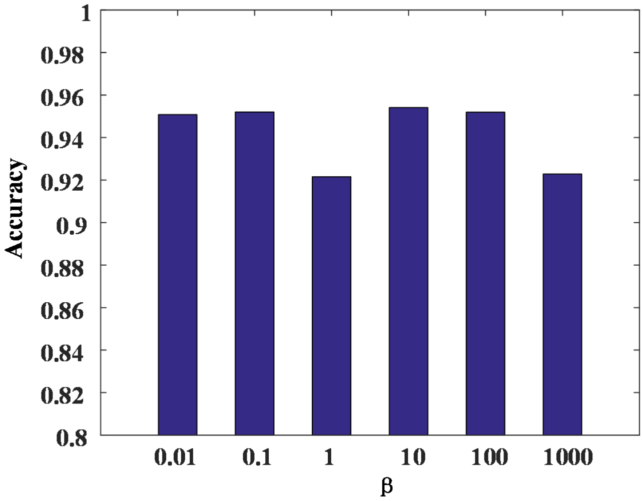


Figure 2: The performance of different $\beta$ according to $\alpha = 1$.

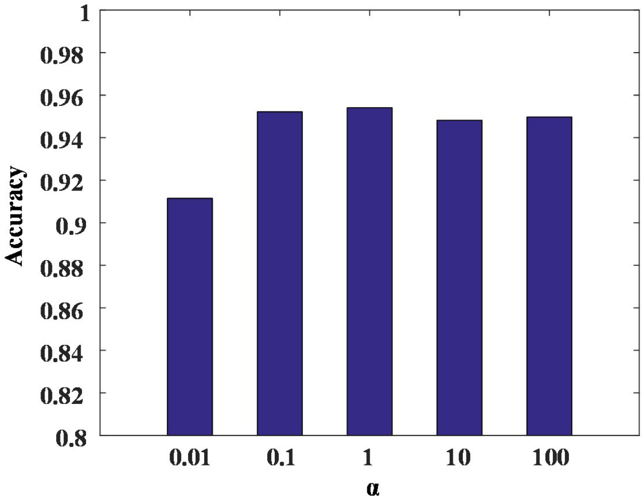


Figure 3: The performance of different $\alpha$ according to $\beta = 10$.

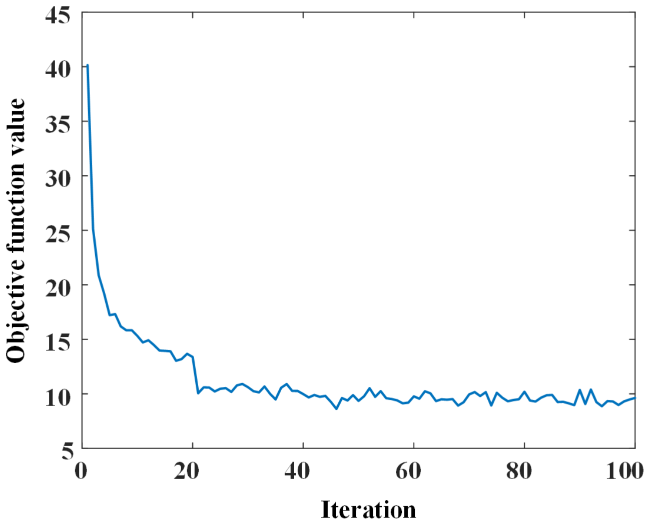


Figure 4: Convergence curve of RRNN on MultiPIE dataset. Here one iteration means the training data input into RRNN once.

classification comes to a stable value after 30 iterations.

## 5. Conclusion

In this paper, we proposed a Recurrent Regression Neural Network (RRNN) to unify two classic face recognition tasks including cross-pose face recognition and video-based face recognition. In RRNN, three basic units are considered to model a potential sequence data. The first unit is the encoder-decoder, which is used to model sequential reconstruction. The second unit is to constrain the globality of the sequence. The final one is to utilize the discriminative label information. By properly choosing the configuration for different tasks, we can benefit from these units. Experimental results strongly indicate our RRNN achieves the best recognition results compared with those state-of-the-art methods.

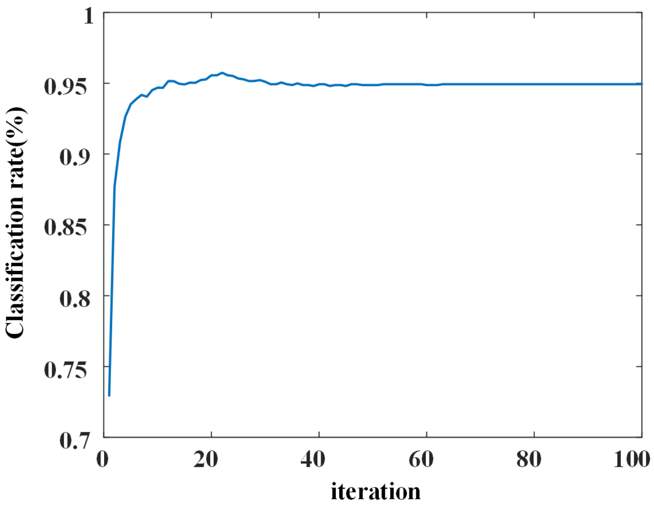


Figure 5: Classification rate versus different number of iterations of RRNN on MultiPIE dataset. Here one iteration means the training data input into RRNN once.

## References

[1] H.-C. Shin, J. H. Park, S.-D. Kim, Combination of warping robust elastic graph matching and kernel-based projection discriminant analysis for face recognition, IEEE Transactions on Multimedia 9 (6) (2007) 1125–1136.

[2] C.-K. Hsieh, S.-H. Lai, Y.-C. Chen, Expression-invariant face recognition with constrained optical flow warping, IEEE Transactions on Multimedia 11 (4) (2009) 600–610.

[3] J. Y. Choi, W. De Neve, K. N. Plataniotis, Y. M. Ro, Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks, IEEE Transactions on Multimedia 13 (1) (2011) 14–28.

[4] G. Liu, P. Li, Low-rank matrix completion in the presence of high coherence, IEEE Transactions on Signal Processing 64 (21) (2016) 5623–5633.

[5] G. Liu, Q. Liu, P. Li, Blessing of dimensionality: Recovering mixture data via dictionary pursuit, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (1) (2017) 47–60.

[6] G. B. Huang, E. Learned-Miller, Labeled faces in the wild: Updates and new reporting procedures, Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep (2014) 14–003.

[7] V. Blanz, T. Vetter, Face recognition based on fitting a 3d morphable model, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1063–1074.

[8] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3d face recognition under expressions, occlusions, and pose variations, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (9) (2013) 2270–2283.

[9] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, M. Rohith, Fully automatic pose-invariant face recognition via 3d pose normalization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 937–944.

[10] M. Kan, S. Shan, H. Chang, X. Chen, Stacked progressive auto-encoders (spae) for face recognition across poses, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1883–1890.

[11] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, IEEE Transactions on Image Processing 19 (6) (2010) 1635–1650.

[12] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (11) (2010) 2106–2112.

[13] B.-C. Chen, C.-S. Chen, W. H. Hsu, Cross-age reference coding for age-invariant face recognition and retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 768–783.

[14] B.-C. Chen, C.-S. Chen, W. H. Hsu, Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset, IEEE Transactions on Multimedia 17 (6) (2015) 804–815.

[15] X. Peng, J. Lu, Z. Yi, R. Yan, Automatic subspace learning via principal coefficients embedding, IEEE Transactions on Cybernetics PP (2016) 1–14.

[16] X. Peng, Z. Yu, Z. Yi, H. Tang, Constructing the l2-graph for robust subspace learning and subspace clustering, IEEE Transactions on Cybernetics 47 (4) (2017) 1053–1066.

[17] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 3444–3451.

[18] J. Gu, X. Yang, S. De Mello, J. Kautz, Dynamic facial analysis: From bayesian filtering to recurrent neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[19] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, S. Shan, Morphable displacement field based image matching for face recognition across pose, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2012, pp. 102–115.

[20] A. Sharma, D. W. Jacobs, Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 593–600.

[21] A. Sharma, A. Kumar, H. Daume III, D. W. Jacobs, Generalized multiview analysis: A discriminative latent space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2160–2167.

[22] J. Lu, G. Wang, W. Deng, P. Moulin, J. Zhou, Multi-manifold deep metric learning for image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 1137–1145.

[23] A. Hadid, M. Pietikäinen, Manifold learning for video-to-video face recognition, in: Biometric ID Management and Multimodal Communication, Springer, 2009, pp. 9–16.

[24] Y.-C. Chen, V. M. Patel, P. J. Phillips, R. Chellappa, Dictionary-based face recognition from video, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2012, pp. 766–779.

[25] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 6645–6649.

[26] A. Graves, J. Schmidhuber, Offline handwriting recognition with multidimensional recurrent neural networks, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2009, pp. 545–552.

[27] Z. Cui, S. Xiao, J. Feng, S. Yan, Recurrently target-attending tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1449–1458.

[28] Z. C. Lipton, J. Berkowitz, C. Elkan, A critical review of recurrent neural networks for sequence learning, arXiv preprint arXiv:1506.00019.

[29] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image and Vision Computing 28 (5) (2010) 807–813.

[30] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.

[31] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition., in: Proceedings of the British Machine Vision Conference (BMVC), Vol. 1, 2015, p. 6.

[32] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.

[33] Y. Bengio, Learning deep architectures for ai, Foundations and trends® in Machine Learning 2 (1) (2009) 1–127.

[34] Z. Lei, D. Yi, S. Z. Li, Learning stacked image descriptor for face recognition, IEEE Transactions on Circuits and Systems for Video Technology 26 (9) (2016) 1685–1696.

[35] O. Yamaguchi, K. Fukui, K.-i. Maeda, Face recognition using temporal image sequence, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 318–323.

[36] T.-K. Kim, J. Kittler, R. Cipolla, Learning discriminative canonical correlations for object recognition with image sets, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2006, pp. 251–262.

[37] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.

[38] R. Wang, X. Chen, Manifold discriminant analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 429–436.

[39] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2567–2573.

[40] Y. Hu, A. S. Mian, R. Owens, Sparse approximated nearest points for image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 121–128.

[41] R. Wang, H. Guo, L. S. Davis, Q. Dai, Covariance discriminative learning: A natural and efficient approach to image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2496–2503.

[42] J. Lu, G. Wang, P. Moulin, Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 329–336.

[43] P. Zhu, L. Zhang, W. Zuo, D. Zhang, From point to set: Extend the learning of distance metrics, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2664–2671.

[44] J. Lu, G. Wang, W. Deng, P. Moulin, Simultaneous feature and dictionary learning for image set based face recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2014, pp. 265–280.

[45] W. Wang, R. Wang, Z. Huang, S. Shan, X. Chen, Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 2048–2057.

**Biography**

**Yang Li** received the B.S. degree in electronic information and science technology from School of Physics and Electronics, Shandong Normal University, China, in 2012, the M.S. degree in electronic and communication engineering from School of Electronic Engineering, Xidian University, China, in 2015. Currently, he is pursuing the Ph.D. degree in information and communication engineering in Southeast University, China. His researches focus on Pattern Recognition and Machine Learning.

**Wenming Zheng** received the B.S. degree in computer science from Fuzhou University, Fujian, China, in 1997, the MS degree in computer science from Huaqiao University, Quanzhou, Fujian, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, Jiangsu, China, in 2004. Since 2004, he has been with the Research Center for Learning Science, Southeast University, Nanjing. He is currently a professor in the Key Laboratory of Child Development and Learning Science of the Ministry of Education, Research Center for Learning Science, Southeast University. His research interests include neural computation, pattern recognition, machine learning, and computer vision. He is a member of the IEEE.

**Zhen Cui** received Ph.D. degrees from Institute of Computing Technology (ICT), Chinese Academy of Sciences in 2014. He was a Research Fellow in the Department of Electrical and Computer Engineering at National University of Singapore (NUS) from 2014 to 2015. He also spent half a year as a Research Assistant on Nanyang Technological University (NTU) from Jun. 2012 to Dec. 2012. Currently he is a Professor of Nanjing University of Science and Technology, China. His research interests cover computer vision, pattern recognition and machine learning, especially focusing on deep learning, manifold learning, sparse coding, face detection/alignment/recognition, object tracking, image super resolution, emotion analysis, etc. He has published several papers in the top conferences NIPS/CVPR/ECCV and some journals of IEEE Transactions. More details can be found in http://aip.seu.edu.cn/zcui/.

**Tong Zhang** received the B.S. degree in Department of Information Science and Technology, Southeast University, China, in 2011, the M.S. degree in Research Center for Learning Science, Southeast University, China, in 2014. Currently, he is pursuing the Ph.D. degree in information and communication engineering in Southeast University, China. His interests include pattern recognition, machine learning and computer vision.