Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses

Meina Kan, Shiguang Shan, Hong Chang, Xilin Chen
Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
{kanmeina, sgshan, changhong, xlchen}@ict.ac.cn

Abstract

Identifying subjects with variations caused by poses is one of the most challenging tasks in face recognition, since the difference in appearances caused by poses may be even larger than the difference due to identity. Inspired by the observation that pose variations change non-linearly but smoothly, we propose to learn pose-robust features by modeling the complex non-linear transform from the non-frontal face images to frontal ones through a deep network in a progressive way, termed as stacked progressive auto-encoders (SPAE). Specifically, each shallow progressive auto-encoder of the stacked network is designed to map the face images at large poses to a virtual view at smaller ones, and meanwhile keep those images already at smaller poses unchanged. Then, stacking multiple these shallow auto-encoders can convert non-frontal face images to frontal ones progressively, which means the pose variations are narrowed down to zero step by step. As a result, the outputs of the topmost hidden layers of the stacked network contain very small pose variations, which can be used as the pose-robust features for face recognition. An additional attractiveness of the proposed method is that no pose estimation is needed for the test images. The proposed method is evaluated on two datasets with pose variations, i.e., MultiPIE and FERET datasets, and the experimental results demonstrate the superiority of our method to the existing works, especially to those 2D ones.

1. Introduction

During the past decades, face recognition has been successfully applied in many areas, such as access control, ID authentication, watch-list surveillance, etc. However, it is still a far way to go for those uncontrolled scenarios due to the large variations caused by the expression, pose, lighting, aging, and so on. Among them, the pose variation is one of the largest challenges, since the facial appearance variations caused by poses are even larger than that caused by identities. To address the pose problem, many promising works have been developed, which can be roughly divided into two categories: 2D techniques and 3D techniques [26].

In most 3D methods, 3D face information, either recovered from the input image or a statistical model learnt in advance is exploited to render a virtual face image at a given pose the same as that of the face image to match. With the virtual view, two face images from different poses can be matched at the same pose. In [8], each face image is represented by the model parameters for the 3D shape and texture, which are estimated by fitting a 3D morphable model to account for the pose variations. In [18], a 3D model is also first estimated for each subject by fitting a 3D generic elastic model to his/her gallery image at frontal pose, and then a group of virtual images at different poses are synthesized, among which the one at the same pose as the probe image is used for the matching. In [3], all face images are compared under the frontal pose by projecting the non-frontal face image onto an aligned 3D face model and then rotating it to render a frontal face image. In [16], a non-frontal face image is transformed to a frontal view by using the morphable displacement field from the 3D face models.

These 3D techniques have achieved favorable performance in many scenarios, even in fully automatic case. However, these methods need 3D data or recovery of 3D face model from 2D images which is still a challenging problem. Besides, automatically fitting a 3D face model to a 2D image is also sensitive to many factors, *e.g.*, illumination, expression, occlusion and so on.

Differently, the 2D methods attempt to handle the pose variations by learning pose-invariant feature or predict the face image under novel target pose without using 3D information. Given an image, some early researchers [7] propose to generate its virtual view at a target pose by learning the transformations between poses. In [11], an eigen-light field model that contains all available pose variations is estimated for each image and used as the pose-invariant feature. In [10], the virtual frontal view of a non-frontal input face image is obtained by applying the learnt locally linear transformations between the non-frontal face images and frontal ones on the densely sampled patches. In [19], the non-

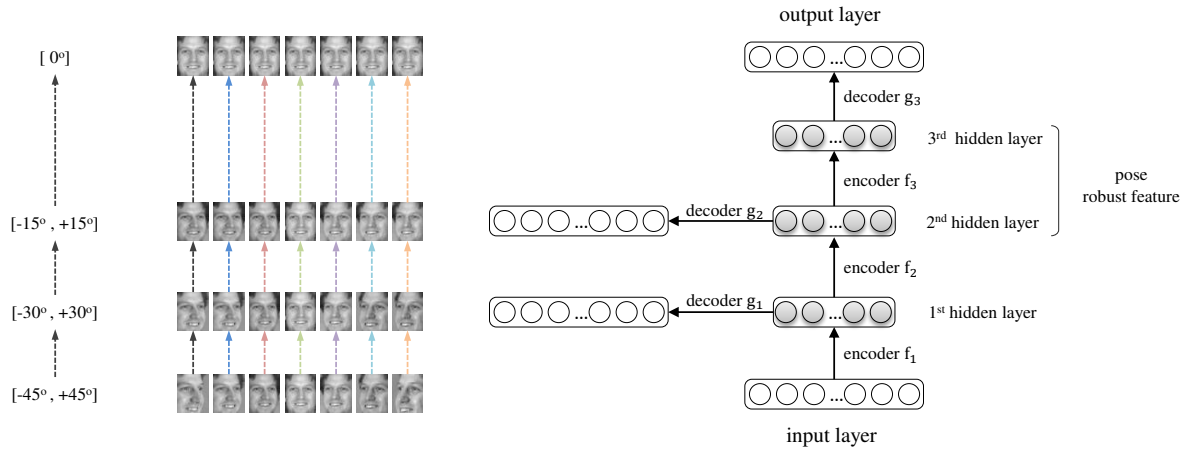


Figure 1. The schema of the proposed Stacked Progressive Auto-Encoders (SPAE) network for pose-robust face recognition. We illustrate an exemplar architecture of the stacked network with $L = 3$ hidden layers, which can deal with poses in yaw rotation within [-45°, +45°]. In training stage of our SPAE, each progressive auto-encoder aims at converting the face images at large poses to a virtual view at a smaller pose (*i.e.*, closer to frontal), and meanwhile keeping the face images with smaller poses unchanged. For instance, for the first progressive AE demonstrated in this figure, only images with yaw rotation larger than 30° are converted to 30°, while other face images with yaw rotation smaller than 30° are mapped to themselves. Such a progressive mode endows each progressive AE a limited goal matching its capacity. In the testing stage, given an image, it is fed into the SPAE network, and the outputs of the topmost hidden layers with very small pose variations are used as the pose-robust features for face recognition.

frontal face image is considered to be generated by a pose-contingent linear transformation of the identity, and the obtained pose-invariant identity subspace is used for recognition. In [2], an alignment strategy referred to as "stack flow" is proposed to discover pose-induced spatial deformities undergone by a face at the patch levels. With this model, a non-frontal face image can be warped to the frontal view incrementally. In [1], the Markov random fields model is used to match two images with local patches as nodes and 2D displacement vectors as their labels. In [9], the correspondence between the frontal and non-frontal face images is obtained by using the dynamic programming based stereo matching algorithm. In [15], an elastic matching method based on Gaussian Mixture Model (GMM) is proposed to match the images at different poses by aligning the patches through the GMM. In [20], a discriminant coupled latent subspace framework is proposed to find a set of projections for different poses such that the projections of the same subject but at different poses are maximally correlated in the latent space. In [28] and [27], deep networks are employed to convert a non-frontal face to a fontal one or a random face, which achieves promising results.

With relative ease of implementation but promising performance, 2D methods are preferred in spite of its slight inferiority than 3D methods in terms of recognition accuracy. However, the pose varies continuously and non-linearly, a single model usually cannot fully characterize all variations. Therefore, most of the 2D methods, *e.g.*, [2] [20], contain pose-specific components to decompose the complex pose variations into piecewise simpler ones. This however means that they generally need to estimate or manually label the pose of a given image, which makes these methods heavily depend on the accuracy of the pose estimation.

In this work, following the basic idea of 2D methods, we propose to extract pose-robust features by learning the complex non-linear transform from the non-frontal face images to frontal ones. Inspired by the observations that pose variations change non-linearly but smoothly, we intend to model the complex non-linear transform from the non-frontal face images to frontal ones through an deep network, considering its impressive ability to handle non-linearity [13] [24]. We specifically resort to deep auto-encoder (DAE) [5] to achieve this goal. However, a direct application of DAE is intractable due to the high complexity of pose variation, especially in case of limited number of training samples. To solve this problem, we propose a progressive deep structure, named Stacked Progressive Auto-Encoders (SPAE), with each shallow progressive AE designed to achieve limited but tractable goal, *i.e.*, part of the global non-linearity. Specifically, as demonstrated in Fig. 1, each shallow AE of our SPAE is designed to convert the input face images at large poses to a virtual view at a smaller pose (*i.e.*, closer to frontal), and meanwhile keep those already with smaller poses unchanged. With such a strategy, we actually enforce the deep network to approximate its eventual goal (frontal pose) layer by layer along the pose manifold (starting from non-frontal face images), as shown in Fig. 2. Or, in other words, our SPAE gradually narrows down the pose variations layer by layer. As a result, the outputs of the topmost hidden layers of the stacked network contain very small pose variations, and thus can be used as the pose-robust features for face recognition as shown in Fig. 1. Please be

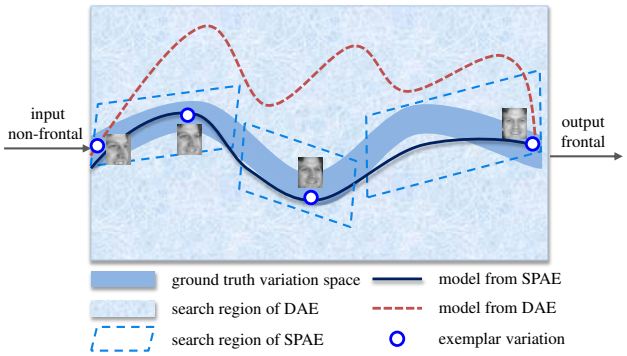


Figure 2. Illustration of non-linear but smooth pose variations, and how SPAE smartly sets each layer of shallow AE a limited but tractable goal.

aware that, in our SPAE, no pose estimation is needed for testing images, which forms a great advantage compared with many previous ones.

The rest of this paper is organized as follows: section 2 details the stacked progressive auto-encoders network; section 3 evaluates it on Multi-PIE and FERET databases, followed by the conclusion in the last section.

2. Stacked Progressive Auto-Encoders (SPAE)

In this section, we first introduce the auto-encoder neural network, and then describe our SPAE deep architecture for pose-robust face recognition.

2.1. Auto-Encoder (AE)

For a shallow auto-encoder neural network [5] which is unsupervised, it is usually compromised of two parts, encoder and decoder [24], with single hidden layer.

The encoder, denoted as $\mathbf{f}$, attempts to map the input $\mathbf{x} \in \mathbf{R}^{d\times 1}$ into the hidden layer representations, denoted as $\mathbf{z} \in \mathbf{R}^{r\times 1}$, in which r is the number of neurons in the hidden layers. Typically, $\mathbf{f}$ consists of a linear transform and a successive nonlinear transform as follows:

$$\mathbf{z} = \mathbf{f}(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{1}$$

where $\mathbf{W} \in \mathbf{R}^{r\times d}$ is the linear transform , $\mathbf{b} \in \mathbf{R}^{r\times 1}$ is the basis and $s(\cdot)$ is the so-called element-wise "activation function", which is usually non-linear, such as sigmoid function $s(x) = \frac{1}{1+e^{-x}}$ or tanh function $s(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$.

The decoder, denoted as $\mathbf{g}$, tries to map the hidden representation $\mathbf{z}$ back to the input $\mathbf{x}$, *i.e.*,

$$\mathbf{x} = \mathbf{g}(\mathbf{z}) = s(\hat{\mathbf{W}}\mathbf{z} + \hat{\mathbf{b}}), \tag{2}$$

with the linear transform $\hat{\mathbf{W}} \in \mathbf{R}^{d\times r}$ and basis $\hat{\mathbf{b}} \in \mathbf{R}^{d\times 1}$.

To optimize the parameters $\mathbf{W}, \mathbf{b}, \hat{\mathbf{W}}$ and $\hat{\mathbf{b}}$, usually the least square error is employed as the cost function:

$$\left[\mathbf{W}^*, \mathbf{b}^*, \hat{\mathbf{W}}^*, \hat{\mathbf{b}}^*\right] = \arg\min_{\mathbf{W},\mathbf{b},\hat{\mathbf{W}},\hat{\mathbf{b}}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{g}(\mathbf{f}(\mathbf{x}_i))\|_2^2, \tag{3}$$

where $\mathbf{x}_i$ represents the i^{th} one of the N training sample. Due to the non-linearity of the activation function, Eq. (3) is difficult to solve, and thus the gradient descent algorithm is commonly employed.

The typical auto-encoder in Eq. (3) tries to reconstruct the input, however if a distinct response rather than same as the input is used as the output of the decoder $\mathbf{g}$, it can be considered as a kind of non-linear regression technique.

2.2. Motivation and basic idea of SPAE

For face recognition across pose, many works attempt to model the non-linear transform from the non-frontal face images to the frontal ones [3] [16]. Most of the successful works exploit the 3D face models to characterize the non-linearity, which however may encounter difficulties when recovering the 3D models. An alternatively preferred strategy is to directly model the non-linear transforms from the non-frontal pose to frontal pose based on the 2D images.

The deep learning technique provides a big opportunity for the above goal in virtue of its great ability for non-linearity, *e.g.*, the auto-encoder can achieve a favorable non-linear regression model. In order to model the complicated transforms from the non-frontal pose to frontal pose, the deep auto-encoders (DAE) network [5] with multiple hidden layers are preferred attributed to its larger capacity compared to the shallow network. A straightforward implementation is to use the non-frontal face images and frontal face images as the input and output of the DAE respectively. This however might be intractable due to the following factors. As shown in Fig. 2, when DAE is used to directly transform the non-frontal face images to the frontal ones, the objective is highly non-linear leading to a larger search region, therefore the DAE is prone to trap into local minima that deviates far from the true one, especially given a relatively small number of training samples.

On the other hand, the pose variations actually change smoothly along a manifold as shown in Fig. 2, which means an easier objective with less non-linearity within a small piece of the manifold. Therefore, in our SPAE, given some "halfway milestones", *e.g.*, those denoted as the blue circles in Fig. 2, the whole challenging objective is decomposed to multiple easier (*i.e.*, less nonlinearity) and thus more tractable phases with smaller search region. It is also in this sense that SPAE has larger probability to avoid some bad local minima.

Specifically, we employ the stacked deep architecture network, in which each shallow part is designed to achieve a limited but tractable goal, *i.e.*, pieces of the global non-linear transform determined by the "halfway milestones". Each shallow AE of this network, called progressive auto-encoders hereafter, attempts to convert the face images at large poses to a virtual view at a smaller pose, but meanwhile keep those images already with smaller poses un-

changed. Then stacking multiple progressive auto-encoders can eventually convert the non-frontal images to frontal ones step by step along the pose variation manifold, as shown in Fig. 2. This stacked deep architecture is termed as Stacked Progressive Auto-encoders (SPAE). The SPAE gradually narrows down the pose variations, and thus induce pose-robust features for further face recognition.

2.3. Formulation of SPAE

For the purpose of clarity, we assume that face poses are divided into $2 \times L + 1$ bins within $[-V, V]$ denoted as $\mathbb{V}$, $-V$ and V are the maximum pose angles face to left and right respectively, and 0° means the frontal pose. For example, in case of $V = 45^\circ, L = 3$, $\mathbb{V}$ might be $\{-45^\circ, -30^\circ, -15^\circ, 0^\circ, +15^\circ, +30^\circ, +45^\circ\}$. To facilitate the presentation, we define an array $\mathbb{P}$ containing the target poses of each progressive auto-encoder in descending order, *e.g.*, $\mathbb{P} = \{+30^\circ, +15^\circ, 0^\circ\}$ excluding the extreme pose which would be not used as the target. $\mathbf{x}_{ij}$ represents the i^{th} sample at pose angle of j with $i \in [1, N]$ and $j \in \mathbb{V}$.

As mentioned, each progressive AE aims at mapping the images at large pose to virtual images at smaller pose, while mapping those already at smaller poses to themselves. For example, the first progressive AE aims at mapping the images at pose larger than 30° to virtual images at 30°, but mapping those at pose smaller than 30° to themselves. In other words, this progressive AE narrows down the pose variations from $[-45^\circ, +45^\circ]$ to $[-30^\circ, +30^\circ]$. Similarly, the second progressive AE is designed to narrow down the pose variations from $[-30^\circ, +30^\circ]$ to $[-15^\circ, +15^\circ]$. Finally, the third progressive AE narrows down the pose variations from $[-15^\circ, +15^\circ]$ to 0°. Therefore, stacking three progressive AEs can convert the non-frontal face images to frontal ones gradually.

Formally, the k^{th} $(k = 1, 2, \cdots, L)$ progressive auto-encoder attempts to convert the images at poses larger than $\mathbb{P}(k)$ to $\mathbb{P}(k)$. The cost function is formulated as bellow:

$$\left[\mathbf{W}_k^*, \mathbf{b}_k^*, \hat{\mathbf{W}}_k^*, \hat{\mathbf{b}}_k^*\right] = \arg\min_{\mathbf{W}_k, \mathbf{b}_k, \hat{\mathbf{W}}_k, \hat{\mathbf{b}}_k} \sum_{i=1}^{N} \sum_{j \in \mathbb{V}} \|\mathbf{x}_{il} - \mathbf{g}_k(\mathbf{f}_k(\mathbf{z}_{ij}^{k-1}))\|_2^2, \quad (4)$$

where $\mathbf{z}_{ij}^{k-1}$ is the representation from the hidden layer (*i.e.*, encoder) of the $(k-1)^{th}$ progressive auto-encoder for the sample $\mathbf{x}_{ij}$, and $\mathbf{z}_{ij}^0 = \mathbf{x}_{ij}$. l is short for l_{ij}^k, which is the target pose that $\mathbf{z}_{ij}^{k-1}$ will be transformed to, calculated as:

$$l_{ij}^k = \begin{cases} -\mathbb{P}(k) & \text{if } l_{ij}^{k-1} < -\mathbb{P}(k) \\ +\mathbb{P}(k) & \text{if } l_{ij}^{k-1} > \mathbb{P}(k) \\ l_{ij}^{k-1} & \text{if } |l_{ij}^{k-1}| \leq \mathbb{P}(k) \end{cases}, \quad (5)$$

with $l_{ij}^0 = j$. $\mathbf{f}_k$ and $\mathbf{g}_k$ are encoder and decoder of the k^{th} progressive auto-encoder. The hidden-layer representation of sample $\mathbf{x}_{ij}$ from the k^{th} progressive auto-encoder is:

$$\mathbf{z}_{ij}^k = \mathbf{f}_k(\mathbf{z}_{ij}^{k-1}) = s\left(\mathbf{W}_k^* \mathbf{z}_{ij}^{k-1} + \mathbf{b}_k^*\right). \quad (6)$$

Each progressive auto-encoder can be optimized using the gradient descent algorithm similarly.

As seen from Eq. (4) and Eq. (5), the pose variations are reduced to $[-\mathbb{P}(k), \mathbb{P}(k)]$ through the k^{th} progressive auto-encoder. Therefore, the pose variations are narrowed down and down gradually by stacking multiple progressive auto-encoders, until no pose variations, *i.e.*, all input face images are converted to virtual frontal face images.

Given a training set with face poses ranged within $[-V, V]$, L progressive auto-encoders are needed in order to convert the images from all poses to the frontal view. After optimizing each progressive auto-encoder, the whole stacked network is comprised of L encoder from all progressive auto-encoders and one decoder from the last one, *i.e.*, $\mathbf{f}_f, \mathbf{f}_2, \cdots, \mathbf{f}_L, \mathbf{g}_L$, as shown in Fig. 1.

After achieving each shallow progressive auto-encoder, the whole network is tuned finely by optimizing all layers jointly as bellow:

$$\left[\mathbf{W}_k^*|_{k=1}^L, \mathbf{b}_k^*|_{k=1}^L, \hat{\mathbf{W}}_L^*, \hat{\mathbf{b}}_L^*\right] = \arg\min_{\mathbf{W}_k|_{k=1}^L, \mathbf{b}_k|_{k=1}^L, \hat{\mathbf{W}}_L, \hat{\mathbf{b}}_L} \sum_{i=1}^{N} \sum_{j \in \mathbb{V}} \|\mathbf{x}_{i,0^\circ} - \mathbf{g}_L(\mathbf{f}_L(\mathbf{f}_{L-1}(\cdots \mathbf{f}_1(\mathbf{x}_{ij}))))\|_2^2. \quad (7)$$

Eq. (7) can be easily solved by employing the gradient descent algorithm, with $\mathbf{f}_f, \mathbf{f}_2, \cdots, \mathbf{f}_L, \mathbf{g}_L$ initialized by the ones learnt from the shallow networks.

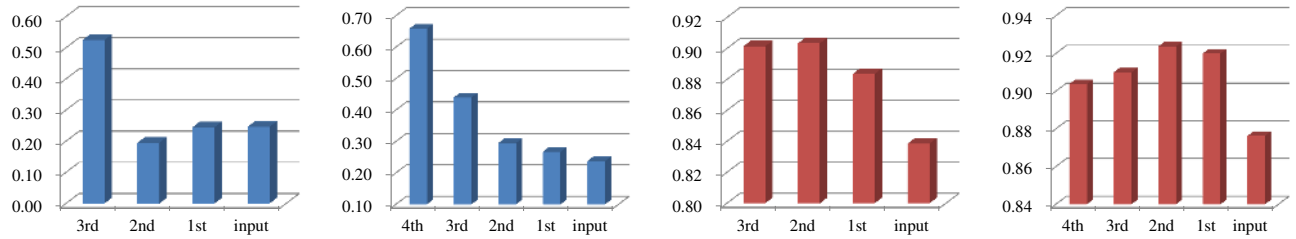
2.4. Pose-robust feature and recognition

As the pose variations are reduced layer by layer, the representation of the topmost layer $\mathbf{f}_L$ should almost have no pose variations, and the representations embedded in the lower layers, *e.g.*, $\mathbf{f}_{L-1}, \mathbf{f}_{L-2}$, are not pose-invariant, but only contain very small pose-variations. As stated in [26], most of the recognition methods are robust to small pose variations. Therefore, in this work, we use the outputs of the few topmost hidden layers as the pose-robust features, denoted as $\mathbf{F}(\mathbf{x})$ and calculated as:

$$\mathbf{F}(\mathbf{x}) = [\mathbf{f}_L(\mathbf{z}^{L-1}); \mathbf{f}_{L-1}(\mathbf{z}^{L-2}); \cdots; \mathbf{f}_{L-k}(\mathbf{z}^{L-k-1})], \quad (8)$$

with $0 \leq k \leq L-1$. As observed in the experiments these representations with small pose variations are more robust than only the pure pose-invariant feature (*i.e.*, outputs of the topmost layer), when combined with Fisher Linear discriminant analysis (FLD) [4] for recognition.

The features $\mathbf{F}(\mathbf{x})$ learnt from the SPAE deep network is unsupervised, so it cannot be expected to be discriminative. Therefore, we further employ the Fisher Linear discriminant (FLD) analysis [4] for supervised dimensionality reduction and the nearest neighbor classifier for recognition.



(a) raw feature on MultiPIE (b) raw feature on FERET (c) raw feature+FLD on MultiPIE (d) raw feature+FLD on FERET

Figure 3. The performance of the learnt feature from each hidden layer of the SPAE network. Here, "1st" means the first hidden layer, and "input" means the original intensity feature. "raw feature" means the feature from each hidden layer is directly used for the recognition which is unsupervised, "raw feature+FLD" means the feature from each hidden layer is followed by an supervised dimensionality reduction method FLD [4]. Three and four hidden layers are included in the stacked network on MultiPIE and FERET datasets respectively.

3. Experiments

In this section, we first evaluate our SPAE w.r.t. different parameters, and then compare it with the state-of-the-art methods on two large scale datasets with pose variations, *i.e.*, the MultiPIE dataset [12] and FERET dataset [17].

3.1. Experimental settings

MultiPIE dataset [12] contains images of 337 people under various poses, illumination and expressions. In this work, the images of all 337 subjects at 7 poses (-45°, -30°, -15°, 0°, +15°, +30°, +45°) with neural expression and frontal illumination are used for the evaluations. This subset is divided into two parts: images from first 200 subjects (subject ID 001 to 200) are used for training, 4,207 images in total. The images from the rest 137 subjects are used for testing, 1,879 images in total. All images are aligned to 40x32 pixels, with five facial landmarks automatically located by using the supervised descent method (SDM) [25], which formulates an automatic setting. Same as [3], the frontal face images from the earliest session for the 137 subjects are used as gallery image (137 in total), and images from the other poses as probe images (1,742 in total).

On the **FERET** dataset [17], the images of all 200 subjects at 9 different poses (bb, bc, bd, be, ba, bf, bg, bh, bi corresponding to -60°, -40°, -25°, -15°, 0°, +15°, +25°, +40°, +60° respectively) are used for the evaluations, one image per subject at each pose. The images of the first 100 subjects are used for training (900 in total), and the images of the rest 100 subjects are used for testing (900 in total). Similarly, the images at the frontal pose are used as the gallery, and the images at the rest poses as the probe images. To compare with the existing methods, semi-automatic setting is employed on this dataset, *i.e.*, all images are aligned to 40x32 by using five manually labeled facial landmarks.

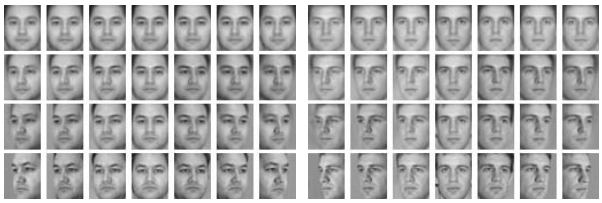
For [14] [21] [22], DAE and SPAE, PCA [23] is applied to for dimensionality reduction, and the dimension of PCA is traversed from 100 to 600 to report the best results.

3.2. Effects of parameters

In SPAE, the number of progressive auto-encoders, *i.e.*, the number of hidden layers, can be determined according to the number of poses (*i.e.*, 3 and 4 on the MultiPIE and FERET datasets respectively), thus leaving two parameters, the number of neurons in each hidden layer, and the number of layers selected to formulate the pose-robust feature.

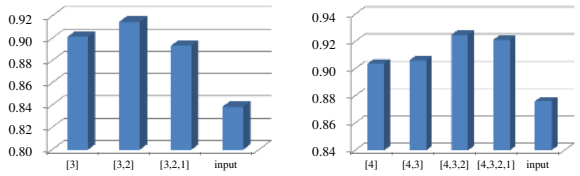
Firstly, we investigate the performance of the features from each hidden layer in the stacked network. The cosine function is used to calculate the similarity, as shown in Fig. 3(a) and Fig. 3(b). As seen, the original intensity features perform badly, due to the large pose variations. Furthermore, the features from the first several hidden layers perform better than the intensity features on the FERET dataset, since pose variations are reduced. However, the features from the first several hidden layers performs only comparable or even worse on MultiPIE dataset, which may be caused by the challenge of the automatic recognition. But, on both datasets, the features from the topmost hidden layer achieve the best performance with a significant improvement, benefited from the fact that all pose variations are removed in this level. An exemplar output of each progressive auto-encoder is shown in Fig. 4. Besides, we also evaluate the decoded face images from each layer and the performance increases layer by layer as expected: The results for the 1st to 4th layer are 22%, 23%, 33% and 59% respectively on FERET, and the results for the 1st to 3rd layer are 22%, 27%, and 61% respectively on MultiPIE.

Although improvements are achieved from the top hidden layers with small pose variations, it is not promising enough for the recognition due to lack of supervised information. Therefore, we apply the supervised FLD to these unsupervised features, as shown in Fig. 3(c) and Fig. 3(d). Benefited from the supervised information, all performances are improved significantly as expected. Another interesting observation is that, the features from the topmost hidden layer with no pose variations do not perform the best any more, while the features from the hidden layers with



(a) (b)

Figure 4. The output from each decoder in the SPAE network for the input images in the bottom row. (a) output of exemplar training images from MultiPIE. (b) output of exemplar testing images from MultiPIE.



(a) MultiPIE (b) FERET

Figure 5. The performance of cumulated features from multiple hidden layers of SPAE network. Here, the numbers in the brackets are the indexes of hidden layers for cumulation, and "input" means the original intensity.

small pose variations (*e.g.*, the second hidden layer on MultiPIE and FERET) perform better. It may seem confusing, but it is actually reasonable. In FLD, a within-class scatter matrix is calculated, which characterises the intra-personal variations including the pose variations. But if all intra-personal pose variations are eliminated completely, it will deteriorates the calculation of within-class scatter matrix leading to a poor performance. Even worse, on the FERET dataset with single sampler per subject for each pose, the FLD would fail to work since only single non-duplicated image is available for each class if the pose variations are removed completely, *i.e.*, zero cost is achieved in Eq. (7). On the other hand, the FLD model is robust to small pose variations as stated in [26], therefore the features from the second hidden layers with pose variations in $[-25°, +25°]$ perform even better than the features from the topmost hidden layer.

We further inspect the performance of cumulated features from multiple hidden layers, *i.e.*, from highest hidden layer to the lower ones, as shown in Fig. 5. The leftmost bar shows the performance of the features from the topmost hidden layer, and the performance improves when the features from one or two more hidden layers are cumulated, because the cumulated features contain pose variations smaller than $25°$ and FLD can easily tolerate them. But if the features with large pose variations, *e.g.*, larger than $25°$, are cumulated, the performance begins to degrade. This demonstrates that the pose-robust feature with small pose variations can achieve a better performance when combined with FLD. Therefore, in the following experiments, the cumulated features with pose variations smaller than $25°$ combined with FLD are used for recognition.

Another important parameter is the number of neurons in the network. More neurons imply a more flexible structure, which can achieve a better mapping from non-frontal face images to the frontal ones. Following the work in [6], we use the same number of neurons in all hidden layers, and explore the performance of the stacked network under different number of neurons, as shown in Table 1. As expected, the more neurons, the better performance. This is easily to understand, since the pose variations are complex and highly non-linear, so a sophisticated network with larger number of neurons is more qualified. From these observations, we use 5000 and 4000 neurons for each hidden layer of SPAE on the MultiPIE and FERET datasets respectively.

Table 1. Results with different number of nodes in the network.

Dataset	number of nodes in each hidden layer				
	1000	2000	3000	4000	5000
MultiPIE	85.7%	88.2%	88.9%	90.9%	91.4%
FERET	89.5%	89.1%	91.5%	92.5%	92.3%

3.3. Comparison with the existing methods

In this section, we compare the SPAE to a few existing methods, which are briefly described as bellow.

"**Blanz03**" [8] proposes to use the parameters of the 3D shape and texture models that are estimated by fitting a statistical 3D model to an image as the representation of it.

"**Asthana11**" [3] proposes a fully automatic 3D pose normalization method, which can synthesize a frontal view of the input face image by aligning an average 3D model to it based on the view-based AAM. This work achieved impressive results under the automatic settings.

MDF [16] proposes to generate a virtual image at the pose of the gallery image for the probe image through the Morphable Displacement Field.

StackFlow [2] warps a non-frontal face image to the fontal one progressively through one or more correspondences between them at the patch level.

CCA [14] aims at projecting images at two different poses to a common space where the cross correlation between them are maximized meaning minimum pose variations.

PLS [21] attempts to project samples from two poses to a common latent subspace, with one pose as regressor and another pose as response.

GMA [22] is a generalized multi-view analysis method attempting to project the images at all poses to a discriminative common space, where pose variations are minimized.

DAE [5] is also evaluated, which has the same structure (*i.e.*, the number of the hidden layers, the number of neurons in each layer) as that of our SPAE network. But differently, it directly converts the non-frontal face images to the frontal

Table 2. Comparison with the existing methods on the MultiPIE dataset under automatic scenario.

Methods		Probe Pose							Pose Estimation
		-45°	-30°	-15°	+15°	+30°	+45°	Average	
3D	Asthana11 [3]	74.1%	91.0%	95.7%	95.7%	89.5%	74.8%	86.8%	Automatic
	MDF [16]	78.7%	94.0%	99.0%	98.7%	92.2%	81.8%	90.7%	
2D	PLS [21]	51.1%	76.9%	88.3%	88.3%	78.5%	56.5%	73.3%	Manually
	CCA [14]	53.3%	74.2%	90.0%	90.0%	85.5%	48.2%	73.5%	
	GMA [22]	75.0%	74.5%	82.7%	92.6%	87.5%	65.2%	79.6%	
	DAE [5]	69.9%	81.2%	91.0%	91.9%	86.5%	74.3%	82.5%	N/A
	SPAE	84.9%	92.6%	96.3%	95.7%	94.3%	84.4%	91.4%	

Table 3. Comparison with the existing methods on the FERET dataset under semi-automatic scenario.

Methods		Probe Pose									Pose Estimation
		bb -60°	bc -40°	bd -25°	be -15°	bf +15°	bg +25°	bh +40°	bi +60°	Average	
3D	Blanz03 [8]	95%	95%	97%	100%	97%	96%	95%	91%	95.8%	Automatic
	MDF [16]	87%	97%	99%	99%	100%	99%	98%	92%	96.4%	
2D	PLS [21]	39%	59%	76%	76%	77%	72%	53%	37%	60.0%	Manually
	CCA [14]	40%	66%	83%	85%	84%	88%	70%	39%	69.4%	
	StackFlow [2]	48%	70%	89%	96%	94%	82%	62%	42%	72.9%	
	DAE [5]	62%	91%	93%	96%	96%	94%	83%	61%	84.5%	N/A
	SPAE	77%	96%	98%	99%	99%	99%	95%	77%	92.5%	

ones, *i.e.*, the input is same as ours, but the output is the frontal face image which is same as the output of the final layer in our SPAE network.

Among the above methods, [8] [3] [16] are 3D-based models which also involve in automatically estimating the pose of the test image, and the rest are 2D-based methods. The 2D methods in [2] [14] [21] [22] assume that the poses of the test images are already known (denoted as "manually"), which implies that they may degenerate when the pose is unavailable. On the contrary, the DAE and our SPAE can work well without any pose estimation (denoted as "N/A").

Firstly, all approaches are evaluated under the automatic scenario on MultiPIE, *i.e.*, the face region and facial landmarks are located automatically. The poses of the test images are unavailable unless specified and the results are shown in Table 2. As seen, CCA and PLS perform the worst since they are fully unsupervised method which are unfavourable for the recognition. Furthermore, GMA performs better benefited from the supervised information. However, they are all inferior to the 3D methods, Asthana11 and MDF, since 3D models exploit more information than 2D models, which may be however unavailable or hard to be collected. DAE has the similar network structure as our SPAE, but it performs worse, since it directly converts the non-frontal images to frontal ones, which cannot guarantee small pose variations within top hidden layers. SPAE achieves the best performance, even compared with the 3D methods. Besides, one great advantage of SPAE is that no pose estimation is needed, even automatic estimation.

Then, all methods are evaluated under the semi-automatic scenario on FERET, with the facial landmarks are manually labeled for face alignment. The result are shown in Table 3. The task on this dataset is easier due to the semi-automatic setting. As a result, the 3D methods perform better than all the 2D methods, including ours. However, SPAE can perform better than the 3D methods when the pose variations are smaller than $45°$. Moreover, our SPAE performs much better than the other 2D methods including DAE.

Overall, our SPAE network can achieve much better performance than 2D methods, and outperform 3D methods on MultiPIE dataset under the automatic setting. This improvement mainly comes from two folds. Firstly, the pose variations from non-frontal poses to the frontal pose are modeled roughly along the intrinsical pose variation manifold, which can guarantee that the pose variations are narrowed down progressively and meanwhile the identities are preserved to facilitate the recognition. Secondly, SPAE does not need to estimate the pose of the test image, avoiding the degeneration from the imperfect estimation of pose.

4. Conclusions and future works

For face recognition across pose, we proposed a stacked progressive auto-encoders network to map non-frontal face images to its frontal view gradually. The global complicated non-linearity from the non-frontal pose to frontal pose is divided into pieces of more tractable ones, which are modeled by multiple shallow progressive auto-encoders respectively. The features from the few topmost layers of the stacked network contain very small pose variations, and thus are further combined with FLD for pose-invariant face recognition. As evaluated, SPAE can effectively reduce the pose variations, and improve the performance of face recognition.

In future, we will extend our SPAE network to deal with illumination, expression, noises and occlusions. Cross-database learning will be also investigated. Besides, we will explore to incorporate the discriminative information into the design of network.

Acknowledgements

This work is partially supported by Natural Science Foundation of China under contracts nos. 61390511, 61222211, 61272319, and 61173065.

References

[1] S. R. Arashloo and J. Kittler. Energy normalization for pose-invariant face recognition based on mrf model image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(6):1274–1280, 2011.

[2] A. B. Ashraf, S. Lucey, and T. Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[3] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 937–944, 2011.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, 1997.

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

[6] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade*, pages 437–478, 2012.

[7] D. Beymer and T. Poggio. Face recognition from one example view. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 500–507, 1995.

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

[9] C. D. Castillo and D. W. Jacobs. Wide-baseline stereo for face recognition with large pose variation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 537–544, 2011.

[10] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose-invariant face recognition. *IEEE Transactions on Image Processing (TIP)*, 16(7):1716–1725, 2007.

[11] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2002.

[12] R. Gross, I. Matthews, J. Cohn, T. kanada, and S. Baker. The cmu multi-pose, illumination, and expression (multi-pie) face database. Technical report, Carnegie Mellon University Robotics Institute. TR-07-08, 2007.

[13] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

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

[15] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3499–3506, 2013.

[16] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *European Conference on Computer Vision (ECCV)*, pages 102–115, 2012.

[17] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(10):1090–1104, 2000.

[18] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(10):1952–1961, 2011.

[19] S. J. Prince, J. H. Elder, J. Warrell, and F. M. Felisberti. Tied factor analysis for face recognition across large pose differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(6):970–984, 2008.

[20] A. Sharma, M. A. Haj, J. Choi, L. S. Davis, and D. W. Jacobs. Robust pose invariant face recognition using coupled latent space discriminant analysis. *Computer Vision and Image Understanding (CVIU)*, pages 1095–1110, 2012.

[21] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 2011.

[22] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[23] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 591, pages 586–591, 1991.

[24] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research (JMLR)*, 9999:3371–3408, 2010.

[25] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[26] X. Zhang and Y. Gao. Face recognition across pose: A review. *Pattern Recognition (PR)*, 42(11):2876–2896, 2009.

[27] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2416–2423, 2013.

[28] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity preserving face space. In *IEEE International Conference on Computer Vision (ICCV)*, pages 113–120, 2013.