Regularized Latent Least Square Regression for Cross Pose Face Recognition

Xinyuan Cai, Chunheng Wang, Baihua Xiao, Xue Chen, Ji Zhou
State Key Laboratory of Management and Control for Complex Systems
Institute of Automations, Chinese Academy of Science, Beijing, China
{xinyuan.cai, chunheng.wang,baihua.xiao,xue.chen,ji.zhou}@ia.ac.cn

Abstract

Pose variation is one of the challenging factors for face recognition. In this paper, we propose a novel cross-pose face recognition method named as Regularized Latent Least Square Regression (RLLSR). The basic assumption is that the images captured under different poses of one person can be viewed as pose-specific transforms of a single ideal object. We treat the observed images as regressor, the ideal object as response, and then formulate this assumption in the least square regression framework, so as to learn the multiple pose-specific transforms. Specifically, we incorporate some prior knowledge as two regularization terms into the least square approach: 1) the smoothness regularization, as the transforms for nearby poses should not differ too much; 2) the local consistency constraint, as the distribution of the latent ideal objects should preserve the geometric structure of the observed image space. We develop an alternating algorithm to simultaneously solve for the ideal objects of the training individuals and a set of pose-specific transforms. The experimental results on the Multi-PIE dataset demonstrate the effectiveness of the proposed method and superiority over the previous methods.

1 Introduction

In the past several decades, face recognition has received a great deal of attentions from the scientific and industrial communities, due to its wide range of applications including access control, security and surveillance, and so on. Although many face recognition approaches [Zhao *et al*., 2003; Wright *et al*., 2009] have reported satisfactory performance, they require accurate alignment and feature correspondence between the face images to be compared. In many real scenarios, the face images maybe captured under various poses, illuminations, or occlusions, which may cause the misalignment problem. According to the survey of face recognition techniques [Zhao *et al*., 2003], the pose variation is identified as one of the most challenging factors.

The difficulty for cross-pose face recognition is that the pose varies in 3D space, while the image captures only 2D appearances. When the pose changes, some visible parts on face may become self-occluded, while some invisible parts may appear. It leads to the special phenomenon that the distance between two different persons with similar pose is smaller than the distance between the same person under different poses. The performance of many popular face recognition methods, such as eigen-face or fisher-face [Belhumeur *et al*., 1997], decreases significantly when confronted with large pose variations.

Many approaches have been proposed to deal with the pose problem [Zhang *et al*., 2009]. A natural approach to solve this problem is to generate the virtual frontal pose images. [Blanz *et al*., 2003] proposed the 3D morphable model (3DMM). For a given 2D face image, they constructed its corresponding 3D model, and did matching in the 3D shape and texture space. However this process is computationally expensive, and relies on manual initialization of several facial points. [Chai *et al*., 2007] proposed the local linear regression method to synthesize a virtual pose image directly in the 2D domain. They assumed that a facial image could be faithfully represented as a linear combination of a set of face images with the same pose, and the coefficients of the linear combination remained roughly constant across poses. [Li *et al*., 2012] reported significantly improved performance by using Ridge regression to estimate the coefficients. [Ying *et al*., 2010] proposed the associate-predict model. They divided the face image into patches, then each patch was associated with a similar patch from a reference set under the same pose. In the prediction step, the associated patch's corresponding patch under the gallery pose was used as proxy for the original patch.

Besides explicitly synthesizing the virtual images, many researchers employ some statistical machine learning methods to find pose-dependent projections, which could project the images into a pose-independent latent subspace. [Li *et al*., 2009] proposed to use Canonical Correlation Analysis (CCA) to learn a pair of projections which make the projected intra-individual features to be maximally correlated in the latent space. [Sharma *et al*., 2011] employed the partial least square regression (PLS) to project samples from two poses to a common latent space. However, these methods considered only two poses at one time, and could not get a common space across all poses. So [Rupnik *et al*., 2010] proposed an extension of CCA for the multi-view problems. [Kan *et al*., 2012] proposed the Multi-view Discriminative Analysis (MvDA) to learn view-specific transforms by maximizing the between-class variation while minimizing the within class variation from all views in a common discriminant space. [Prince *et al*., 2008] proposed a generative model, named Tied Factor Analysis (TFA), for cross pose face recognition. They assumed that the observed images could be generated by a pose-specific linear transformation

of identity variables in the presence of Gaussian noise. They used the EM algorithms to learn the model parameters from a set of training images in different known poses. In recognition step, they computed the probability that the probe and gallery share the same identity vectors.

The common goal of these existing approaches is to build a bridge between the observed image space and the pose free representation space. Since the images of one person are captured under different poses of the same face, they are highly related to each other. Similar to [Prince *et al*., 2008], it is reasonable to assume that the images taken under different poses can be viewed as pose-specific transforms of a single ideal object. In real applications, there exist some noises, such as illumination, expression *et al*., so this assumption may not hold exactly. We relax this hard assumption, and try to find a set of pose-specific transforms, which make the projected images of the same person as close as possible to the single ideal object. We treat the observed images as regressor, the latent ideal objects as response, and then formulate this relaxed assumption in a compact least square regression framework. Compared with previous methods, the advantages of our method are that: **1)** we formulate the problem of cross multi-poses face recognition in a compact framework; **2)** two regularization terms can be naturally incorporated into the regression based approach. The first is the smoothness regularization, which enforces the transforms for nearby poses to be close to each other. The second is the local consistency regularization, which enforces that the distribution of the latent ideal objects should preserve the geometric structure of the observed image space; **3)** the final optimization problem can be solved efficiently by an alternating algorithm.

The remainder of this paper is organized as follows. In Section 2, we give a brief review of some related works. In Section 3, we present the formulation of our proposed method, and develop an alternating algorithm to solve the optimization problem efficiently. In Section 4, we evaluate the proposed method on the Multi-PIE dataset [Gross *et al*., 2007], followed by the conclusions.

2 Related Works

2.1 Canonical Correlation Analysis and Partial Least Square Regression

Canonical Correlation Analysis (CCA) and Partial Least Square Regression (PLS) are two classic statistical techniques to find linear relations between two multidimensional variables. Formally, let $X^1 = [x_1^1, x_2^1, \dots, x_n^1]$ and $X^2 = [x_1^2, x_2^2, \dots, x_n^2]$ represent the sets of facial images from two different poses respectively, where $x_i^1 \in R^M, x_i^2 \in R^N$. And the samples in X^1 and X^2 are coupled by identities. The goal of both CCA and PLS is to find two linear transformations $W_1 \in R^{M\times d}, W_2 \in R^{N\times d}$ for X^1 and X^2 respectively, so that the projected images of the same person are similar to each other in the projected latent subspace. The definition of similarity varies with the methods, so they are different in their objective functions. CCA tries to maximize the correlation between the projected images of the same person:

$$\max_{W_1,W_2} \left(\mathrm{corr}\left[W_1^T X^1, W_2^T X^2\right]\right) \quad (1)$$
$$s.t.\ \|W_1\|_F=1,\ \|W_2\|_F=1$$

where $\mathrm{corr}(a,b) = \mathrm{cov}(a,b)/(\mathrm{var}(a)\times \mathrm{var}(b))$. (2)

PLS tries to maximize the covariance between the projected images of the same person:

$$\max_{W_1,W_2} \left(\mathrm{cov}\left[W_1^T X^1, W_2^T X^2\right]\right) \quad (3)$$
$$s.t.\ \|W_1\|_F=1,\ \|W_2\|_F=1$$

CCA can be naturally extended to deal with multi-view problems, which is called Multi-view CCA [Rupnik *et al*,.2010]. The goal of MCCA is to find a set of linear transformations $\{W_i\}(i=1,\dots,p)$, one for each pose, so that the sum of all pair-wise correlations is maximized:

$$\max_{W_1,\dots,W_p} \sum_{i<j} \mathrm{corr}\left[W_i^T X^i, W_j^T X^j\right] \quad (4)$$
$$s.t.\ \|W_i\|_F = 1, i = 1, \dots, p$$

2.2 Tied Factor Analysis

Tied Factor Analysis (TFA) [Prince *et al*., 2008] is a generative model to describe pose variations of facial images. They introduce two spaces: the observed space and the identity space. The observed space means the raw pixel feature or a transformed space after simple pose-independent transformations. In this space, the locations of pose variant and identity variant facial images are mixed altogether, which makes the separation of identity from pose variations infeasible. To effectively separate the identity from pose variations, an ideal identity space is introduced. The underlying assumption is that the face images of a person across different poses can be generated from a common latent variable in the identity space. The relationship between these two spaces can be described as:

$$x_{ij}^k = F_k h_i + M_k + E_{ij}^k \quad (5)$$

where x_{ij}^k representes the jth image of individual *i* under pose *k*. h_i is the latent identity vector for individual *i*. F_k and M_k are the linear transformation and offset respectively for pose *k*. E_{ij}^k is a zero-mean Gaussian noise with unknown diagonal covariance matrix Σ_k. More formally, the model can be written in terms of conditional probabilities:

$$\Pr(x_{ij}^k | h_i) = Gaussian(F_k h_i + M_k, \Sigma_k) \quad (6)$$
$$\Pr(h_i) = Gaussian(0, I) \quad (7)$$

From a set of training images in different known poses, the identity vectors and the pose-specific transformations can be estimated by maximizing the joint likelihood of the observed image data *x* and its associated identity *h*.

$$\max_{\Theta,h} \Pr(x, h|\Theta) \quad (8)$$

where $\Theta = \{F_{1,\dots,k}, M_{1,\dots,k}, \Sigma_{1,\dots,k}\}$ is the model parameters. They use the EM algorithm [Dempster *et a*l., 1977] to learn the model parameters, which is prone to local minima and computationally intensive. In the recognition phase, they compute the probability that the probe face and the gallery face are generated by exactly the same value of the identity vector under the linear transformation scheme, and the final decision is made through maximum a posteriori mechanism by choosing the gallery image which corresponds to the maximum probability.

3 Regularized Latent Least Square Regression

3.1 Problem Formulation and Learning Model

Formally, we assume the pose can be discretized into *p* different bins. Let $\{X^k\}$, k=1,…,p denote a set of training images, where $X^k = \{x_{ij}^k | i = 1, \dots, c, j = 1, \dots, n_i^k\}$ denotes the samples in pose *k*, $x_{ij}^k \in R^d$ is the *j*th image of individual *i* from

pose k in d dimensions, n_i^k is the number of images for individual i under pose k, and c is the number of individuals.

Since the images are captured from different poses of the same person, it is reasonable to assume that there are some common features across all poses. Both of PLS [Sharma *et al.,* 2011] and CCA [Li *et al*., 2009] approaches try to find a set of projections, which make the projected images of the same person from different poses similar to each other in the latent pose-free space. Similar to the TFA [Prince *et al*., 2008], we assume the idealized version, which means that the images of the same person from different poses can be projected to the same identity vector in the latent space. So we reformulate the Eq(5) as:

$$h_i = F_k^{-1}x_{ij}^k - F_k^{-1}M_k - F_k^{-1}E_{ij}^k \quad (9)$$

(9) can be written in a generalized form:

$$h_i = f(w_k, x_{ij}^k) + m_k + \varepsilon_{ij}^k \quad (10)$$

where $h_i \in R^m$ is the idealized identity vector of individual i; $f(w_k, x_{ij}^k)$ is the mapping function for pose k, and $w_k \in R^{d\times m}$ is the parameter. $m_k \in R^m$, and ε_{ij}^k are the offset and noise term respectively for pose k. For simplicity, we use the linear mapping function:

$$f(w_k, x_{ij}^k) = w_k^T x_{ij}^k \quad (11)$$

Our goal is to obtain the parameters $\{w_k, m_k\}_{k=1}^p$ of a set of pose-dependent (but identity-independent) mapping functions, so that in the recognition phase, we can get the identity vectors h_g and h_p of the gallery x_g and probe x_p. But as the noise term ε exists, we just could get the estimated identity vectors $\hat{h}_g$ and $\hat{h}_p$:

$$\hat{h}_g = f(w_{pose(g)}, x_g) + m_{pose(g)} \quad (12)$$

$$\hat{h}_p = f(w_{pose(p)}, x_p) + m_{pose(p)} \quad (13)$$

For the training set, the estimated identity vector should be as close as possible to the idealized identity vector (as shown in Fig 1). We formulate the above motivation in the least square sense as:

$$\min_{h_{1,..,c}, w_{1,...,p}, m_{1,...,p}} \sum_{k=1}^{p} \frac{1}{N^k} \sum_{i=1}^{c} \sum_{j=1}^{n_i^k} \left\| h_i - f(w_k, x_{ij}^k) - m_k \right\|_2^2 \quad (14)$$

$$s.t.\ \|h_i\|_2^2 = 1 \quad i=1,\ldots,c$$

The formulation (14) can be interpreted as a least square regression problem, if we treat the observed images as regressor, the parameters of mapping functions as regression matrixes, and the identity vectors as response. But the idealized identity vectors cannot be observed directly, so we call the formulation (14) as the Latent Least Square Regression.

Let $X_i^k = [x_{i1}^k, \ldots, x_{in_i^k}^k] \in R^{d\times n_i^k}$ collect the images of individual i under pose k, and $X^k = [X_1^k, X_2^k, \ldots, X_c^k] \in R^{d\times N^k}$ collect all the individuals' images under pose k.

$$F_i^k = [\underbrace{0,\ldots,0}_{\sum_{t=1}^{i-1} n_t^k}, \underbrace{1,\ldots,1}_{n_i^k}, \underbrace{0,\ldots,0}_{\sum_{t=i}^{c} n_t^k}] \in R^{1\times N^k} \quad (15)$$

$$A^k = [F_1^k; F_2^k; \ldots; F_c^k] \in R^{c\times N^k} \quad (16)$$

$$H = [h_1, h_2, \ldots, h_c] \in R^{m\times c} \quad (17)$$

$$e_{N^k} = [1,1,\ldots,1]^T \in R^{N^k\times 1} \quad (18)$$

Then the objective function (14) can be rewritten as:

$$\min_{h_{1,..,c}, w_{1,...,p}, m_{1,...,p}} \sum_{k=1}^{p} \frac{1}{N_k} \left\| HA^k - f(w_k, X^k) - m_k e_{N^k}^T \right\|_F^2 \quad (19)$$

$$s.t.\ \|h_i\|_2^2 = 1 \quad i=1,\ldots,c$$

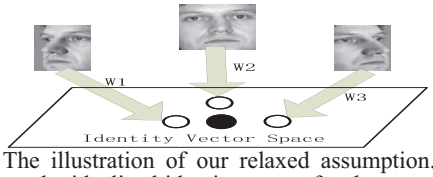

Figure 1: The illustration of our relaxed assumption. The solid circle denotes the idealized identity vector for the example person. The hollow circles denote the estimated identity vectors for the different pose images. For the images of the same person, the estimated identity vectors should be as close as possible to the idealized identity vector.

where $\|.\|_F$ stands for the Frobenius norm of matrix.

However, using the least square approach in (19) is not enough for finding a faithful mapping, because the least square approach may lead to the problem of over fitting and thus has poor generalization ability. In order to deal with this problem, we impose some prior knowledge as regularizations into the objective function to limit the solution space and improve the generalization performance.

1) Smoothness regularization. Though we want to project the facial images in distinct poses by different mapping functions, the facial images under nearby poses don't differ too much. Hence, the mapping functions should vary as smoothly as possible when the pose varies. This smoothness regularization can also tolerate small pose estimation error, which will be validated in the experimental part. The smoothness regularization can be written as:

$$R_{smoothness} = \sum_{k=1}^{p-1} \|w_k - w_{k+1}\|_F^2 \quad (20)$$

2) Local consistency regularization. As revealed in many previous studies [Roweis *et al*., 2000; He *et al*., 2005], locality information is an important clue in manifold learning. In most manifold learning methods, researchers try to find a subspace that best preserves the local structure of the observed data. In this work, we could regard the idealized identity vectors as the low-dimensional embeddings of the observed facial images. So the distribution of the identity vectors in the identity space should preserve the local geometric structure of the observed image space, which means if two persons are similar in the observed image space, their identity vectors should be close to each other. We achieve this objective by minimizing the following energy function:

$$R_{consistency} = \sum_{i=1}^{c}\sum_{j=1}^{c} \left\| h_i - h_j \right\|^2 S_{ij} = tr(HBH^T) \quad (21)$$

where h_i, h_j is the identity vectors for individual i and j; S_{ij} is the similarity between individual i and j, $S = \{S_{ij}\}$, $B = D - S$ is the Laplacian matrix, and D is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. We select one image $\hat{x}$ of each person under the same pose (e.g., frontal view), and define S_{ij} as:

$$S_{ij} = \begin{cases} e^{\frac{-\|\hat{x}_i - \hat{x}_j\|_2^2}{\sigma_i \sigma_j}} & \text{if } \hat{x}_i \in N_k(\hat{x}_j) \text{or } \hat{x}_j \in N_k(\hat{x}_i) \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

where $N_k(\hat{x}_i)$ denotes the set of k nearst-neighbors of $\hat{x}_i$. σ_i is the distance from $\hat{x}_i$ to its qth nearest neighbor, and q is a local scale factor. This way of defining similarity matrix is called local scaling [Manor *et al*., 2005], and it controls the decreasing speed of S_{ij} with the distance between $\hat{x}_i$ and $\hat{x}_j$. The local parameter k and q influence the relative contribution of local structure preserving in the observed space.

In addition, in order to control the complexity of mapping function, an extra normalization term $\sum_{k=1}^{p} \|w_k\|_F^2$ needs to

be included. To sum up, the proposed Regularized Latent Least Square Regression model is formulated as follows:

$$\{h_{1,..,c}, w_{1,...,p}, m_{1,...,p}\} = \text{argmin } L \quad (23)$$
$$s.t. \ \ \|h_i\|_2^2 = 1 \quad i=1,\ldots,c$$

where $L = \sum_{k=1}^{p} \frac{1}{N^k} \left\| HA^k - f(w_k, X^k) - m_k e_{N^k}^T \right\|_F^2 + \lambda \sum_{k=1}^{p-1} \|W_k - W_{k+1}\|_F^2 + \eta tr(HBH^T) + \gamma \sum_{k=1}^{p} \|w_k\|_F^2$ (24)

and λ, η, γ are nonnegative trade-off parameters.

3.2 Solving the Optimization Model

Here we develop an alternating method to solve the optimization problem (23). We decompose the original problem into two key subproblems as following:

Subproblem 1: Given the identity vectors $\{h_i\}_{i=1}^c$, the mapping functions' parameters $\{w_k, m_k\}_{k=1}^p$ can be obtained by solving the following optimization problem:

$$\{w_{1,...,p}, m_{1,...,p}\} = \text{argmin } g(\{w_{1,...,p}, m_{1,...,p}\}) \quad (25)$$

where

$$g(\{w_{1,...,p}, m_{1,...,p}\}) = \sum_{k=1}^{p} \frac{1}{N^k} \left\| HA^k - f(w_k, X^k) - m_k e_{N^k}^T \right\|_F^2 + \lambda \sum_{k=1}^{p-1} \|w_k - w_{k+1}\|_F^2 + \gamma \sum_{k=1}^{p} \|w_k\|_F^2 \quad (26)$$

A straightforward way to minimize (25) is the gradient descent approach. According to the matrix theory, the derivative of $g(\{w_{1,...,p}, m_{1,...,p}\})$ respect to $\{w_k, m_k\}(k = 1, \ldots, p)$ can be computed as:

$$\partial g/\partial w_k = -2X^k/N^k \times \left(HA^k - w_k^T X^k - m_k e_{N^k}^T\right)^T + 2\lambda \times \delta(k > 1)(w_{k-1} - w_k) + 2\lambda \times \delta(k < p)(w_k - w_{k+1}) + 2\gamma \times w_k \quad (27)$$

$$\partial g/\partial m_k = -2/N^k \times \left(HA^k - w_k^T X^k - m_k e_{N^k}^T\right) \times e_{N^k} \quad (28)$$

where $\delta(x)$ is the indicator function, which returns 1 if x is true, and 0 otherwise. After obtaining the gradient, the parameters $\{w_k, m_k\}(k = 1, \ldots, p)$ can be updated by (29) until convergence

$$w_k = w_k - \alpha \, \partial g/\partial w_k \qquad m_k = m_k - \alpha \, \partial g/\partial m_k \quad (29)$$

Subproblem 2: Given the mapping functions' parameters $\{w_k, m_k\}_{k=1}^p$, the identity vectors $\{h_i\}_{i=1}^c$ can be obtained from the following optimization problem:

$$\{h_i\}_{i=1}^c = \text{argmin } g(h_1, \ldots, h_c) \quad (30)$$
$$s.t. \ \ \|h_i\|_2^2 = 1 \quad i=1,\ldots,c$$

where $g(h_1, \ldots, h_c) = \sum_{k=1}^{p} \frac{1}{N^k} \left\| HA^k - f(w_k, X^k) - m_k e_{N^k}^T \right\|_F^2 + \eta \, tr(HBH^T)$ (31)

Naturally, we consider the derivative of $g(h_1, \ldots, h_c)$ with respect to H, and take it to be zero. This gives us the analytical solution as follows:

$$\partial g/\partial H = 0$$
$$\Rightarrow \sum_{k=1}^{p} \frac{2}{N^k} (HA^k - f(w_k, X^k) - m_k e_{N^k}^T) A^{kT} + 2\eta \times HB = 0$$
$$\Rightarrow H = \left(\sum_{k=1}^{p} \frac{1}{N^k} (w_k^T X^k + m_k e_{N^k}^T) A^{kT}\right)\left(\sum_{k=1}^{p} \frac{1}{N^k} A^k A^{kT} + \eta B\right)^{-1} \quad (32)$$

On the basis of above two subproblems, we develop an alternating method to solve the original optimization problem (23). Algorithm 1 lists the steps of RLLSR algorithm.

Algorithm 1: **R**egularized **L**atent **L**east **S**quare **R**egression

Input: Training set $\{X^k\}$, k=1,…,*p*, where $X^k = \{x_{ij}^k | i = 1, \ldots, c, j = 1, \ldots, n_{ik}\}$ are the samples in pose k, and x_{ij}^k is the *j*th image of individual *i* under pose *k*; the positive tradeoff parameters λ, η, γ; maximum number of iterations *T*; step size α.

Output: The parameters $\{w_k, m_k\}_{k=1}^p$ of pose-dependent mapping functions and the identity vectors $H=\{h_i\}_{i=1}^c$ for the training individuals.

```
Select one frontal pose image of each person from the training set, and construct the similarity matrix S according to (22).
Randomly initialize {h_i}_{i=1}^c, {w_k, m_k}_{k=1}^p, and normalize ||h_i||_2^2 = 1 (i=1,…,c); Iter =1.
while Iter<T do
  Compute the objective value L of (24)
  t=1;
  while (t<T) do
    for k=1:p
      Compute the gradient of ∂g/∂w_k (27), ∂g/∂m_k(28)
      Compute w̃_k and m̃_k (k=1,…,p) by
        w̃_k = w_k − α ∂g/∂w_k  m̃_k = m_k − α ∂g/∂m_k
    end for
    delta = Σ_{k=1}^p ||w̃_k − w_k||_F^2 + Σ_{k=1}^p ||m̃_k − m_k||_F^2
    w_k = w̃_k,  m_k = m̃_k.(k=1,…,p);  t=t+1;
    if (delta<1e-5)
      Break;
    end if
  end while
  Compute H according to (32), and normalize ||h_i||_2^2 = 1
  Compute the new objective value L_new of (24)
  if(|L_new − L|<1e-5)
    Objective function converges. Break.
  end if
  Iter= Iter+1;
end while
```

4 Experiments

Like any other learning based methods [Li *et al*., 2009; Kan *et al*., 2012], we require training data to learn the model parameters. We assume access to a training dataset that has multiple images of each person under different known poses. In the training phase, we obtain the pose-dependent mapping functions. At testing time, we assume the pose of the gallery and probe images are known, and use the corresponding mapping functions to estimate the identity vectors; then we adopt the cosine distance to measure the similarity of the estimated identity vectors, and the nearest neighbor classifier is chosen to do the recognition task.

4.1 Dataset and Experimental Settings

We conducted experiments on CMU Multi-PIE [Gross *et al*., 2007] dataset. It contains 337 subjects, recorded during four sessions under various poses, illumination and facial expressions. The interval pose angle is about 15^o. Following [Kan *et al*., 2012], we select 6 images with neutral expression and no flush illuminations of each subject under seven poses ($-45^o, -30^o, -15^o, 0^o, 15^o, 30^o, 45^o$), so a subset about $(6 \times 7 \times 337)$ 14154 images is selected as the evaluation dataset. For the experiments, this dataset is split into a training set, which is used to obtain the pose-dependent mapping functions' parameters $\{w_k, m_k\}(k = 1, \ldots, p)$ and a testing set, which is used to perform the closed-set identification. The first 231 subjects are used for training, and the next 106 subjects for testing. We used the publicly available labeled locations of facial points provided by [Sharma *et al*., 2011] to crop the face regions, and then normalize to 32×32.

Images are turned into gray-scale and direct intensity values mapped between 0 and 1 are used as features. Thus the length of the pixel feature is 1024. Some normalized examples are shown in Fig 2.

Parameters selection is a key issue. The trade-off parameters $\{\lambda, \eta, \gamma\}$ of our model are chosen from {1e-4, 1e-3,…, 1,…,1e3,1e4} and the dimension *m* of identity vector is chosen from {50,100,150,…,1000} by performing cross validation on the training data. In our experiment, we finally set $\lambda = 0.1, \eta = 0.01, \gamma$=0.01, and *m*=200. We implement the proposed method in MATLAB, and the cropped images and source code are available upon request.

Figure 2: Examples of normalized faces of Multi-PIE

4.2 Effectiveness of the Regularization Terms

We evaluate the performance of the proposed model with and without the regularization terms, which are the two objective functions (23) and (33) respectively:

$$\{h_{1,\ldots,I}, w_{1,\ldots,p}, m_{1,\ldots,p}\} = \arg\min \hat{L} \quad (33)$$
$$s.t. \ \ \|h_i\|_2^2 = 1 \quad i=1,\ldots,c$$

where

$$\hat{L} = \sum_{k=1}^{p} \frac{1}{N^k} \left\| H_k - f(w_k, X^k) - m_k e_{N^k}^T \right\|_F^2 + \gamma \sum_{k=1}^{p} \|w_k\|_F^2 \quad (34)$$

Similar to other methods [Li *et al*., 2009; Prince *et al*., 2008; Kan *et al*., 2012], we use the images from one pose as gallery, images from another pose as probe, and then perform closed-set identification. This yields a correct recognition rate for all pairs of poses. The results are shown in Table 1&2 respectively. The last row and column of these two tables denote the average recognition rate.

The comparison between Table 1&2 clearly highlights the improvement offered by using the smoothness and local consistency regularizations. We especially observe a significant improvement (near or over **10**%) for gallery and probe with large pose difference (the result in bold). The average improvement of 42 pairs of gallery and probe setting is **6.9**%. The reason for the improvement can be interpreted from the Eq(27). Without the smoothness regularization, we just use the samples from the *k*th pose to compute the gradient $\partial g / \partial w_k$. While with the smoothness regularization, we consider the relationship between the *k*th pose and its nearby ((*k*-1)th and (*k*+1)th) poses, thus the samples from nearby poses can also contribute to computing the mapping function for pose *k*. And the local consistency constraint works as a term to utilize the structure of the observed data, which helps to improve the generalization.

4.3 Experiment with Inaccurate Probe Pose

In section 4.2, we assume that the pose angles of gallery and probe are known. In order to evaluate the robustness to inaccurate pose estimation, we conduct experiments with unknown probe pose. We assume that the gallery poses are known, while the probe poses are estimated. We evaluate the performance when the difference between the estimated pose and the real pose are $\pm 15^o, \pm 30^o, \pm 45^o$. For example, if the real pose is α, and the estimated pose is β, we use the mapping function for pose β to map the probe images into the identity space followed by nearest neighbor matching.

Fig 3 shows the results when we set the gallery pose to frontal. It can be seen that when the error of pose estimation is within $\pm 15^o$, the performance declination is small. But as the pose estimation error increases, the performance decreases significantly. So it is difficult to recognize faces with big error in pose estimation. In Table 3, we shown the results for all gallery and probe poses when the pose error is within $\pm 15^o$. By comparison with Table 1, it indicates that besides the difference between the real pose and estimated pose, the gap between the gallery and real probe pose is another factor that influences the performance. Table 4 shows the mean accuracy of 42 pairs of gallery and probe pose settings under different pose estimation errors. When the pose estimation error is $\pm 15^o$, the performance decreases 5.5%, but it is still much better than the average performance shown in Table 2. So in some sense, it also demonstrates the effectiveness of the regularization terms.

Gallery \ Probe	-45º	-30º	-15º	0º	15º	30º	45º	Avg
-45º	100	98.3	98.1	**94.8**	**93.1**	**93.6**	**91.7**	94.9
-30º	99.1	100	99.8	99.2	**98.1**	**96.7**	**91.7**	97.4
-15º	99.5	100	100	100	99.5	97.3	**93.1**	98.2
0º	97.3	99.5	100	100	100	99.5	96.9	98.9
15º	**92.8**	97.6	99.5	100	100	99.8	98.6	98.1
30º	**90.7**	**94.3**	**95.3**	98.0	99.5	100	99.0	96.1
45º	**91.5**	**87.6**	**91.7**	**94.7**	98.6	98.6	100	93.8
Avg	95.2	96.2	97.4	97.8	98.1	97.6	95.2	**96.8**

Table 1: Rank-1 recognition rate of our model with the regularization terms.

Gallery \ Probe	-45º	-30º	-15º	0º	15º	30º	45º	Avg
-45º	100	97.2	93.6	**85.7**	**77.7**	**80.0**	**73.3**	84.6
-30º	96.2	100	99.8	93.2	**88.8**	**83.5**	**77.5**	89.8
-15º	95.3	99.8	100	99.2	95.6	91.0	**82.2**	93.9
0º	89.9	98.1	100	100	99.7	99.5	91.0	95.7
15º	**83.8**	93.6	98.0	98.9	100	95.4	91.7	94.0
30º	**75.9**	**86.5**	**84.4**	90.3	97.0	100	92.6	87.8
45º	**74.8**	**75.8**	**77.2**	**83.0**	90.6	100	100	83.2
Avg	86.0	91.8	92.1	91.7	91.6	91.0	84.7	**89.9**

Table 2: Rank-1 recognition rate of our model without the regularization terms.

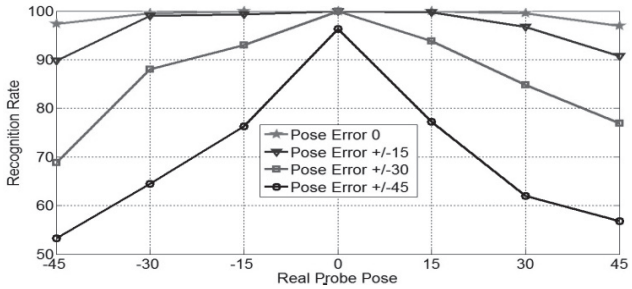

Figure 3: Experimental results with inaccurate probe poses and frontal gallery pose.

4.4 Comparison of Different Approaches

We compare the proposed method (RLLSR) with following algorithms:

- Fisherface model (FLDA) [Belhumeur *et al.*,1997]. It is a baseline method, and it is directly applied on the training set regardless of the pose variations.
- Tied factor Analysis (TFA) [Prince *et al*., 2008]. We use the code provided by the author to evaluate its performance on the Multi-PIE dataset.

Gallery \ Probe	-45°	-30°	-15°	0°	15°	30°	45°	Avg
-45°	100	98.1	96.4	85.7	85.8	83.9	78.0	88.2
-30°	97.9	100	99.8	93.2	94.0	88.4	77.1	92.3
-15°	97.3	99.8	100	99.2	97.3	92.1	80.8	94.4
0°	89.8	99.1	99.4	100	99.7	96.8	90.7	95.9
15°	82.6	93.2	96.1	98.9	100	97.9	94.5	94.0
30°	76.9	87.4	87.6	90.3	95.7	100	97.1	90.0
45°	71.9	78.2	76.5	83.0	95.3	97.4	100	84.3
Avg	86.1	92.6	92.6	94.1	94.6	92.7	86.4	**91.3**

Table 3: Experimental results for all gallery and probe poses when the probe pose error is within $\pm 15^{o}$.

Pose error	**0°**	**±15°**	**±30°**	**±45°**
Mean accuracy	96.8	91.3	80.4	63.0

Table 4: The mean accuracy of all gallery and probe poses under different estimation errors of probe pose.

- Partial Least Square Regression (PLS) [Sharma *et al*., 2011] and Canonical Correlation Analysis (CCA) [Li *et al.*, 2009]. Both of them are pairwise method, which means they just consider two poses at the training and testing phases.
- Multi-view CCA [Rupnik *et al.*, 2010] and Multi-view Discriminative Analysis (MvDA) [Kan *et al*., 2012]. Both of them learn the pose-dependent projections for multi-poses simultaneously. We re-implement these two methods, and get similar results as reported in their original papers.

Fig 4 illustrates the performance comparisons of different approaches under frontal gallery faces. Besides that we also evaluate the rank-one recognition rate under all possible pairs of gallery and probe pose settings (as shown in Table 5~7). And these results are averaged as mean accuracy (as shown in Table 8). From these results, we can observe that:
(1)The performance of TFA, MCCA, MvDA and RLLSR is much better than the pair-wise methods (PLS,CCA), which demonstrates the effectiveness of considering multi-poses simultaneously.
(2)The FLDA uses the multi-poses images, but it learns just one projection. It achieves lower accuracy than the method using different projections for different pose images (e.g TFA, MCCA, MvDA and RLLSR). So it is feasible and effective to project different pose images respectively into a common space.
(3)All methods achieve comparable accuracy, when the difference between the gallery pose and probe pose is in $\pm 15^{o}$. However as the pose difference increases, the performance of PLS, CCA, TFA and MCCA degrade significantly. Compared with PLS, TFA, and MvDA, our proposed RLLSR gains 24%, 11% and 2.6% improvement respectively when the pose difference between gallery and probe is larger than 45^{o}.
(4)The proposed RLLSR achieves a mean accuracy of 96.8%, which is significantly better than other methods. The MvDA approach employs the fisher criterion to find a discriminant common space, while the proposed RLLSR doesn't apply any discriminative information. It demonstrates that the proposed generative model with effective regularization terms can obtain better relationships between the observed image space and the latent common space, than MvDA and other methods.

5 Conclusions and Future Works

In this paper, we have developed an approach, named as Regularized Latent Least Square Regression, to deal with the cross pose face recognition problem. We assume that the images of one person captured under different poses could be mapped into a single ideal point in the latent pose-free space. The RLLSR framework provides such a way to formulate this assumption that it can integrate some regularization technology to improve the generalization performance efficiently. Comparative experiments indicated that the proposed method results in high accuracy and robustness for the cross pose face recognition problem. Our future work involves the extension to nonlinear mapping functions, and the exploration of using different holistic or local features at fiducial points.

Gallery \ Probe	-45°	-30°	-15°	0°	15°	30°	45°	Avg
-45°	100	96.3	92.1	86.2	78.3	80.4	72.0	84.2
-30°	95.1	100	98.8	94.2	87.6	84.5	78.2	89.7
-15°	93.5	97.8	100	99.4	95.5	90.0	83.2	93.2
0°	90.0	98.2	99.4	100	99.4	97.5	91.0	95.9
15°	84.5	93.5	98.0	99.0	100	96.4	92.5	94.0
30°	76.0	87.1	85.4	89.5	98.0	100	95.6	88.6
45°	75.0	76.8	78.2	83.5	90.6	100	100	84.0
Avg	85.7	91.6	92.0	91.7	91.6	91.5	85.4	**90.0**

Table 5: Results of TFA for all possible gallery-probe pairs on Multi-PIE.

Gallery \ Probe	-45°	-30°	-15°	0°	15°	30°	45°	Avg
-45°	100	94.2	85.2	78.5	69.3	61.3	60.0	74.8
-30°	93.1	100	92.4	90.1	84.3	70.0	62.3	82.0
-15°	84.5	94.3	100	92.5	83.6	75.0	64.3	82.4
0°	81.0	86.2	93.0	100	94.5	82.5	79.0	86.0
15°	67.5	72.0	80.0	93.0	100	93.4	78.5	80.7
30°	61.0	64.1	75.3	80.5	95.0	100	94.3	78.4
45°	62.3	66.4	70.2	76.5	87.4	94.2	100	76.2
Avg	74.9	79.5	82.7	85.2	85.7	79.4	73.1	**80.1**

Table 6: Results of PLS for all possible gallery-probe pairs on Multi-PIE

Gallery \ Probe	-45°	-30°	-15°	0°	15°	30°	45°	Avg
-45°	100	99.2	95.5	79.5	90.0	92.0	91.5	91.3
-30°	99.6	100	100	95.1	95.0	97.0	94.5	96.9
-15°	92.5	99.3	100	100	99.6	98.8	93.0	97.2
0°	80.0	94.2	100	100	100	97.5	86.5	93.0
15°	88.0	94.0	99.0	100	100	99.2	95.5	96.0
30°	89.5	97.1	97.3	98.0	100	100	98.3	96.7
45°	91.3	95.0	90.6	88.0	94.5	99.2	100	93.1
Avg	90.2	96.5	97.1	93.4	96.5	97.3	93.2	**94.8**

Table 7: Result of MvDA for all possible gallery-probe pairs on Multi-PIE

Method	FLDA	PLS	CCA	TFA	MCCA	MvDA	RLLSR
Mean Accuracy	86.2	78.0	83.2	90.0	92.0	94.8	**96.8**

Table 8: Mean accuracy of different methods for all possible gallery-probe pairs on Multi-PIE.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under grants No.61172103, No.61271429, No.60933010, and Tencent Research Institute of Beijing.

References

[Blanz *et al*., 2003] Blanz, V., Vetter, T. Face recognition based on fitting a 3D morphable model. IEEE Transaction on PAMI, 25, 1063–1074,2003

[Belhumeur et al., 1997] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces:Recognition using class specific linear projection. IEEE Transactions on PAMI. 19, 711 – 720, 1997.

[Chai *et al.*, 2007] X.Chai, S.Shan, X. Chen, W.Gao. Locally linear regression for pose-invariant face recognition. IEEE Trans on Image Process. 16 (7), 1716–1725, 2007

[Dempster *et al*., 1977] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Soc. B, 39, 1-38, 1977.

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

[Hastie *et al.*, 2001] Hastie.T, Tibshirani.R, and Friedman.J. The elements of statistical Learning. Springer, 2001.

[He *et al*., 2005] X.He, S.Yan, Y.Hu, P.Niyogi, and H.Zhang. Face recognition using Laplacianfaces. IEEE Transactions on PAMI, 27(3),328-340,2005

[Kim *et al.*, 2006] Kim, T.-K., Kittler, J., Cipolla, R.: Learning Discriminative Canonical Correlations for Object Recognition with Image Sets. In ECCV 2006.

[Kan *et al*., 2012] M. Kan, S. Shan, H. Zhang, S. Lao X. Chen. Multi-view Discriminate Analysis. In ECCV 2012.

[Li *et al*., 2012] A. Li, S. Shan, and W. Gao. Coupled bias-variance tradeoff for cross-pose face recognition. IEEE Trans on Image Process, 21(1):305–15, 2012.

[Li *et al.*, 2009] A. Li, S. Shan, X. Chen and W. Gao, Maximizing Intra-individual Correlations for Face Recognition Across Pose Differences. In CVPR 2009.

[Ma et al., 2007] Ma, Y., Lao, S., Takikawa, E., Kawade, M.: Discriminate analysis in correlation similarity measure space. In ICML 2007.

[Manor et al., 2005] L.Z.Manor, P.Perona. Self-tuning spectral clustering. In NIPS, 2005.

[Prince et al., 2008] S.J.D. Prince, J. Warrell, J.H. Elder, F.M. Felisberti, Tied factor analysis for face recognition across large pose differences, IEEE Transaction on PAMI. 30 (6) 970–984, 2008

[Rupnik *et al.*, 2010] Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. In: SiKDD 2010

[Rosipal *et al*., 2006] Rosipal, R., Kramer, N. Overview and recent advances in partial least squares. Subspace Latent Struct. Feat. Select., 34-51. 2006.

[Roweis et al., 2000] S.Roweis, L.Saul. Nonlinear dimensionality reduction by local linear embedding. Science,5500: 2323-2326, 2000

[Sharma *et al*.,2011] Sharma, A., Jacobs, D.W.: Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In CVPR 2011

[Wright *et al*., 2009] J.Wright, A.Y.Yang, A.Ganesh, S.Sastry, and Y.Ma. Robust face recognition via sparse representation. IEEE Transactions on PAMI, 31(2), 210-227, 2009

[Ying *et al.*, 2010] Q. Ying, X. Tang and J. Sun. An Associate-Predict Model for Face Recognition. In CVPR 2010

[Zhao *et al*., 2003] W.Zhao, R.Chellappa, P.J.Phillips, and A.Rosenfeld. Face recognition: A literature survey. ACM Computing Surveys, 35(4), 399-458, 2003.

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