Pose-Dependent Low-Rank Embedding for Head Pose Estimation

Handong Zhao[†], Zhengming Ding[†] and Yun Fu[†‡]
[†]Department of Electrical and Computer Engineering, Northeastern University, Boston, USA, 02115
[‡]College of Computer and Information Science, Northeastern University, Boston, USA, 02115
{hdzhao,allanding,yunfu}@ece.neu.edu

Abstract

Head pose estimation via embedding model has been demonstrated its effectiveness from the recent works. However, most of the previous methods only focus on manifold relationship among poses, while overlook the underlying global structure among subjects and poses. To build a robust and effective head pose estimator, we propose a novel Pose-dependent Low-Rank Embedding (**PLRE**) method, which is designed to exploit a discriminative subspace to keep within-pose samples close while between-pose samples far away. Specifically, low-rank embedding is employed under the multi-task framework, where each subject can be naturally considered as one task. Then, two novel terms are incorporated to align multiple tasks to pursue a better pose-dependent embedding. One is the cross-task alignment term, aiming to constrain each low-rank coefficient to share the similar structure. The other is pose-dependent graph regularizer, which is developed to capture manifold structure of same pose cross different subjects. Experiments on databases CMU-PIE, MIT-CBCL, and extended YaleB with different levels of random noise are conducted and six embedding model based baselines are compared. The consistent superior results demonstrate the effectiveness of our proposed method.

Introduction

Head pose estimation is an integral component in computer vision system, which has a wide range of applications, such as face recognition, person/face identification and human-machine interaction. Although there are extensive works studying head pose estimation from 2D images, it is still far away from mature. The challenges are induced by changing illuminations, various facial expressions, and subject variability. A generic algorithm for head pose estimation has to be robust to such factors, e.g., occlusion, noise, lighting and perspective distortion, which make it a challenging issue.

Recently, a number of algorithms have been proposed to address head pose estimation problem, and a good survey can be referred to (Murphy-Chutorian and Trivedi 2009). In general, these existing methods can be categorized into the following groups: template model (Kwong and Gong 2002), regression model (Haj, Gonzàlez, and Davis 2012; Geng and Xia 2014), embedding model (Wang and Song 2014), active appearance model (Edwards et al. 1998; He, Sigal, and Sclaroff 2014) and geometric model (Wang and Sung 2007; Fanelli, Gall, and Van Gool 2011). Among them, embedding model attracts lots of attention recently because of the high accuracy and good generalizability (Fu and Huang 2006; Balasubramanian, Ye, and Panchanathan 2007; Wang et al. 2008; BenAbdelkader 2010; Wang and Song 2014). However, it is worth noticing that all the previous embedding based methods fail to consider the underlying global structure among subjects and poses, which results in vulnerability of noises or corruptions.

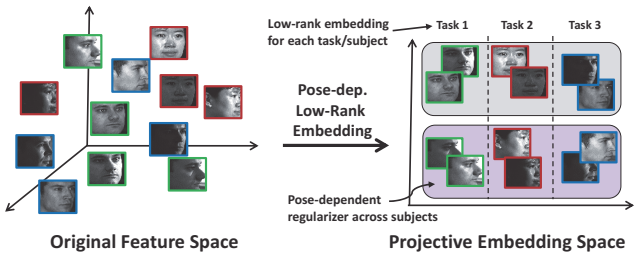


Figure 1: Conceptual illustration of the proposed PLRE method. Head pose images distribute arbitrarily in the original feature space due to the noise and/or some large corruptions, e.g., illumination, etc. Here, for simplicity, we take three subjects as an example marked by different colors. Through the proposed PLRE method, low-rank embedding is employed on each subject under the multi-task framework, where each subject is a task. Regions with different background colors (gray and purple) denote pose-dependent regularizers across subjects, which couple data with same pose but from different subjects. PLRE is optimized jointly to seek the discriminative embedding space, with which the testing head pose is estimated.

To better handle head pose estimation problem, we get inspiration from the fact that data from different poses within the same subject should lie in separable subspaces, which is independent of illuminations, expression, lighting condition, etc. Correspondingly, low-rank representation (Liu et al. 2013; Wang and Fu 2015; Shao, Kit, and Fu 2014; Li and Fu 2014) is a promising way to uncover the pose-specific subspaces within each subject. Low-rank representation manages to find the lowest rank representation

of data, therefore, it can discover the intrinsic multiple subspace structure. Recently, low-rank representation has been incorporated with subspace learning, whose goal is to achieve a robust representation (Ding, Shao, and Fu 2014; Wang, Xu, and Leng 2013). All these methods are designed to uncover the multiple subject-specific subspaces, where they treat each subject as one independent low-rank structure. However, for head pose estimation problem, exploring pose relationship across subjects is the gist. In other words, mining pose-dependent multiple structures plays a key role.

In this paper, we develop a novel Pose-dependent Low-Rank Embedding (**PLRE**) method, which incorporates the cross-task alignment and pose-dependent regularizer into multi-task learning framework, as shown in Figure 1. The core of PLRE is to learn a discriminative subspace, where different poses within the same subject intend to share the similar low-rank structure through a cross-task regularizer, and same poses from different subjects are coupled tightly. Therefore, the learned subspace is more pose-dependent and robust to noise, illumination, subject variations, etc. The main contributions are summarized as follows:

- Low-rank embedding is incorporated into each subject to discover the global structure of pose data under multi-task learning framework. Each subject correlates with one task. This practice reduces the subject-dominant influence in pose data.
- Cross-task alignment term is introduced to impose all the low-rank representations to a similar structure. That is, the low-rank representation in each task is guided by other tasks to achieve the structured coefficients.
- A novel pose-dependent regularizer is designed to couple data with same pose but from different subjects. Thus a more discriminative subspace is learned to keep within-pose samples close while between-pose sample far away.

Pose-dependent Low-Rank Embedding

Problem Formulation

Given a training dataset $X = [X_1, \ldots, X_i, \ldots, X_K]$ with K subjects and $X_i \in \mathbb{R}^{d\times n}$, where n is the sample size of X_i, d is the dimension of the features. Each subject contains several kinds of poses. Thus, $X \in \mathbb{R}^{d\times N}$, where $N = K\times n$ denotes the total number of training samples. The high similarity across different poses within one subject would definitely destroy the performance in head pose estimation. Regarding each subject as a task, our goal is to seek a discriminative subspace $P \in \mathbb{R}^{d\times m}$ (m is the reduced dimensionality) to better separate within-task (within-subject) samples across different poses, whilst keep those between-task (between-subject) samples in the same pose close to each other. Therefore, it is the key to find the intrinsic subspaces of different poses in the same task. We employ low-rank constraint in order to find the multiple subspace structure within each task. Specifically, we introduce low-rank constraint on the low-dimensional projected data. Since we learn low-rank representation for each task separately, for this paper, we formulate the multi-task embedding problem as:

$$\begin{aligned} &\min_{P,Z_i,E_i} \sum_{i=1}^{K} (\|Z_i\|_* + \lambda\|E_i\|_{2,1}), \\ &\text{s.t. } P^{\mathrm{T}}X_i = P^{\mathrm{T}}X_iZ_i + E_i, \\ &\qquad P^{\mathrm{T}}P = I_m,\ i = 1, \ldots, K \end{aligned} \tag{1}$$

where λ is the balancing parameter, and $Z_i \in \mathbb{R}^{n\times n}$ is the low-rank representation of the i-th subject data X_i. $\|Z_i\|_*$ is the trace norm of Z_i, which denotes the sum of singular values of matrix Z_i and is a good surrogate of $\mathrm{rank}(Z_i)$ (Candès et al. 2011). Following the previous works (Zhao and Fu 2015; Ding and Fu 2014), $\ell_{2,1}$-norm is proposed to model the error $E_i \in \mathbb{R}^{m\times n}$ [1], since $\ell_{2,1}$-norm can better detect the corrupted samples caused by partial occlusion, extreme illuminations, or others. $I_m \in \mathbb{R}^{m\times m}$ denotes the identity matrix. The constraint $P^{\mathrm{T}}P = I_m$ enforces the projection matrix P to be orthogonal to avoid the trivial solution and eliminate redundancy. Note that when data sampling is sufficient, X itself works well as the dictionary (Liu et al. 2013). In the following part, two novel regularizers are introduced to the current formulation to achieve a discriminative pose-dependent subspace.

Cross-task alignment regularizer: Inspired by the fact that different subjects should share the similar structure in term of the pose information, we specifically design a cross-task alignment regularizer. For simplicity, we assume the size of same pose in different subjects is equivalent in the training stage. Therefore, each Z_i should have the similar structure, since the pose labels for different subjects are the same. Our target is to align different tasks by constraining the learned Z_i to share the similar structure. To this end, we design a cross-task alignment term $\sum_{j=1,j\neq i}^{K} \|Z_i - Z_j\|_{\mathrm{F}}^2$ in our objective to constrain each Z_i as:

$$\begin{aligned} &\min_{P,Z_i} \sum_{i=1}^{K} (\|Z_i\|_* + \lambda\|E_i\|_{2,1} + \gamma \sum_{j=1,j\neq i}^{K} \|Z_i - Z_j\|_{\mathrm{F}}^2), \\ &\text{s.t. } P^{\mathrm{T}}X_i = P^{\mathrm{T}}X_iZ_i + E_i, \\ &\qquad P^{\mathrm{T}}P = I_m,\ i = 1, \ldots, K, \end{aligned} \tag{2}$$

where γ is the trade-off parameter.

Remark 1: The dual regularization term $\|Z_i - Z_j\|_{\mathrm{F}}^2,\ (i \neq j)$ imposes any two different subjects to share the analogous structure. Equivalently, all the low-rank embedding coefficients are similar when the model reaches optimization.

Remark 2: $\|Z_i - Z_j\|_{\mathrm{F}}^2$ is not oversimplification. In the training stage, due to the accessibility to pose/indentity label, we supervise sort the data samples in order without loss of generality. While in the testing stage, there is no need to know any label information.

Pose-dependent graph regularizer: To make the subspace P more discriminative, we design to couple the same

[1] $\ell_{2,1}$-norm is designed to model sample-specified error, which is defined as $\|E_i\|_{2,1} = \sum_{k=1}^{m} \sqrt{\sum_{j=1}^{n} |E_i|_{kj}^2}$.

pose data from different subjects, which is under the assumption that pose image can be recovered by a linear combination of those images from same pose across different subjects. Consequently, we construct a binary graph G with pose-dependent property, whose element is defined as:

$$G(k,l) = \begin{cases} 1, & k,l \in \text{same pose but different subjects} \\ 0, & \text{otherwise} \end{cases}$$

Then, we add a graph regularizer to align the same pose data from different subjects, and the final objective function is rewritten as:

$$\begin{aligned} \min_{P,Z_i,E_i} & \sum_{i=1}^{K}(\|Z_i\|_* + \lambda\|E_i\|_{2,1} + \gamma \sum_{j=1,j\neq i}^{K} \|Z_i - Z_j\|_{\mathrm{F}}^2) \\ & + \beta \mathrm{tr}(P^{\mathrm{T}} X L X^{\mathrm{T}} P), \\ & \text{s.t. } P^{\mathrm{T}} X_i = P^{\mathrm{T}} X_i Z_i + E_i, \\ & \quad P^{\mathrm{T}} P = I_m,\ i = 1, \ldots, K, \end{aligned} \tag{3}$$

where β is the weight of this regulation term. L is the Laplacian matrix of G, defined as $L = D - G$. Here D is a diagonal matrix, its on-diagonal elements is computed as the corresponding row sums of G. And $\mathrm{tr}(\cdot)$ is the operator to calculate the trace of matrix.

Remark 3: This graph term captures the manifold information of pose data. We connect the same pose data and cut the others, this practice can be seen as the "1-NN" version of graph. Note that the binary graph G can be easily extended to connect the most $\mathcal{K}$ similar poses.

Optimization

Obviously the proposed objective is hard to find the global optimal solution for P, Z_i, E_i jointly. Moreover, the trace norm on Z_i is non-smooth. To solve the problem (3), we employ Augmented Lagrangian Multiplier (ALM) (Lin, Chen, and Ma 2009). For ease of the optimization, we introduce auxiliary variables J_i to relax the problem (3) as:

$$\begin{aligned} \min_{P,Z_i,J_i,E_i} & \sum_{i=1}^{K}\big(\|J_i\|_* + \lambda\|E_i\|_{2,1} + \gamma \sum_{j=1,j\neq i}^{K} \|Z_i - Z_j\|_{\mathrm{F}}^2\big) \\ & + \beta \mathrm{tr}(P^{\mathrm{T}} X L X^{\mathrm{T}} P), \\ & \text{s.t. } P^{\mathrm{T}} X_i = P^{\mathrm{T}} X_i Z_i + E_i, J_i = Z_i, \\ & \quad P^{\mathrm{T}} P = I_m,\ i = 1, \ldots, K, \end{aligned} \tag{4}$$

whose Lagrangian function $\mathcal{L}$ is written as follows:

$$\begin{aligned} \mathcal{L} = & \sum_{i=1}^{K}\big(\|J_i\|_* + \lambda\|E_i\|_{2,1} + \gamma \sum_{j=1,j\neq i}^{K} \|Z_i - Z_j\|_{\mathrm{F}}^2 \\ & + \langle \Pi_i, Z_i - J_i\rangle + \langle \Lambda_i, P^{\mathrm{T}} X_i - P^{\mathrm{T}} X_i Z_i - E_i\rangle \\ & + \frac{\mu}{2}(\|P^{\mathrm{T}} X_i - P^{\mathrm{T}} X_i Z_i - E_i\|_{\mathrm{F}}^2 + \|Z_i - J_i\|_{\mathrm{F}}^2)\big) \\ & + \beta \mathrm{tr}(P^{\mathrm{T}} X L X^{\mathrm{T}} P), \end{aligned} \tag{5}$$

where Λ_i and Π_i are Lagrange multipliers and $\mu > 0$ is a penalty parameter, and $\langle \cdot \rangle$ is the inner product of two matrixes. Then by applying ALM, the variables are optimized independently in an iterative manner. Specifically, J_i, Z_i, E_i and P are updated in $(t+1)$-th iteration as follows:

Update $\mathbf{J_{i,t+1}}$: Fix $Z_{i,t}$, $E_{i,t}$, P_t and solve

$$\arg\min_{J_i} \frac{1}{\mu_t}\|J_i\|_* + \frac{1}{2}\|J_i - (Z_{i,t} + \Pi_{i,t}/\mu_t)\|_{\mathrm{F}}^2. \tag{6}$$

Here trace norm $\|\cdot\|_*$ is difficult to be optimized due to its non-smooth property. However, it is the convex envelope of rank function over the unit ball of spectral norm, which can be recovered by several recently proposed solutions effectively. Here, Singular Value Thresholding (SVT) proposed by Cai *et al*. (Cai, Candès, and Shen 2010) is applied.

Update $\mathbf{Z_{i,t+1}}$: Fix $J_{i,t}$, $E_{i,t}$, P_t and solve the following problem,

$$\begin{aligned} Z_{i,t+1} = & \big(2\gamma(K-1)I_n + \mu_t(R_{i,t}^{\mathrm{T}} R_{i,t} + I_n)\big)^{-1} \\ & \big(2\gamma \sum_{j=1,j\neq i}^{K} Z_{j,t} + R_{i,t}^{\mathrm{T}}(\Lambda_{i,t} + \mu_t R_{i,t} - E_{i,t}) - \Pi_{i,t} + J_{i,t+1}\big), \end{aligned} \tag{7}$$

where $R_{i,t} = P_t^{\mathrm{T}} X_i$ and I_n is the identity matrix with n-dimension.

Update $\mathbf{E_{i,t+1}}$: Fix $J_{i,t}$, $Z_{i,t}$, P_t and solve the following problem,

$$E_{i,t+1} = \arg\min_{E_i} \frac{\lambda}{\mu_t}\|E_i\|_{2,1} + \frac{1}{2}\|E_i - \widehat{E}_{i,t}\|_{\mathrm{F}}^2. \tag{8}$$

where $\widehat{E}_{i,t} = P_t^{\mathrm{T}} X_i - P_t^{\mathrm{T}} X_i Z_{i,t+1} + \Lambda_{i,t}/\mu_t$. This problem can be solved by the off-the-shelf solver (Yang et al. 2009).

Update $\mathbf{P_{t+1}}$: Fix $J_{i,t}$, $Z_{i,t}$, $E_{i,t}$ and solve the following problem,

$$P_{t+1} = (\sum_{i=1}^{K} \mu_t U_{i,t} + 2\beta X L X^{\mathrm{T}})^{-1} \sum_{i=1}^{K} V_{i,t}, \tag{9}$$

where $U_{i,t} = (X_i - X_i Z_{i,t+1})(X_i - X_i Z_{i,t+1})^{\mathrm{T}}$ and $V_{i,t} = (X_i - X_i Z_{i,t+1})(\mu_t E_{i,t+1}^{\mathrm{T}} - \Lambda_i^{\mathrm{T}})$. Then, we enforce P_{t+1} to be orthogonal via $P_{t+1} \leftarrow$ *orthogonalize*(P_{t+1}). The details of the solution are outlined in **Algorithm 1**.

Algorithm 1 *Solving* ***PLRE*** *using ALM*

Input: Training sample X, parameter λ, γ, β

Initialize: $J_{i,0} = Z_{i,0} = E_{i,0} = \Lambda_{i,0} = \Pi_{i,0} = 0$, t = 0, $\mu_0 = 10^{-3}, \rho = 1.2, \epsilon = 10^{-3}, \mu_{\max} = 10^6$.

Output: Z_i, J_i, E_i, P

while not converged **do**

1. Fix the others and update $J_{i,t+1}$ using Eq. (6)
2. Fix the others and update $Z_{i,t+1}$ using Eq. (7)
3. Fix the others and update $E_{i,t+1}$ using Eq. (8).
4. Fix the others and update P_{t+1} using Eq. (9), $P_{t+1} \leftarrow$ *orthogonalize*(P_{t+1}).
5. Update multipliers $\Lambda_{i,t+1}, \Pi_{i,t+1}$ by $\Lambda_{i,t+1} = \Lambda_{i,t} + \mu_t(P_{i,t+1}^{\mathrm{T}} X_i - P_{i,t+1}^{\mathrm{T}} X_i Z_{i,t+1} - E_{i,t+1})$ $\Pi_{i,t+1} = \Pi_{i,t} + \mu_t(Z_{i,t+1} - J_{i,t+1})$,
6. Update parameter μ_{t+1} by $\mu_{t+1} = \min(\rho\mu_t, \mu_{\max})$
7. Check the convergence condition by $\|P_{i,t+1}^{\mathrm{T}} X_i - P_{i,t+1}^{\mathrm{T}} X_i Z_{i,t+1} - E_{i,t+1}\|_\infty < \epsilon$, $\|Z_{i,t+1} - J_{i,t+1}\|_\infty < \epsilon$.
8. $t = t + 1$.

end while

Once we have the optimal solution P^*, both training samples and test samples are projected onto P^*, and then nearest neighbor (NN) classifier is utilized to predict the label of testing samples. We outline the procedures in **Algorithm 2**.

Algorithm 2 PLRE *for Head Pose Estimation*

Input: Training data X with the corresponding label L_X testing data Y.

Output: Predicted label vector L_Y for testing data.

1. Normalize the training samples x_i by $x_i = x_i/\|x_i\|$.
2. Solve problem (3) by **Algorithm 1** and get the optimal projection matrix P^*.
3. Project X and Y to P^* via $\tilde{X} = P^{*\mathrm{T}}X$, $\tilde{Y} = P^{*\mathrm{T}}Y$.
4. Predict the label vector L_Y of $\tilde{Y}$ using nearest neighbor (NN) classifier with cosine distance.

Complexity Analysis

In this subsection, we analyze time complexity of our proposed PLRE. The most time-consuming parts in Algorithm 1 are Steps 1, 2 and 4. The dimensions of the important variables are listed as follows: $X_i \in \mathbb{R}^{d\times n}$, $Z_i \in \mathbb{R}^{n\times n}$, $J_i \in \mathbb{R}^{n\times n}$ and $P \in \mathbb{R}^{d\times m}$. Accordingly, in the first step, the SVT operator needs singular value decomposition of matrices of $O(n^3)$. In Step 2 and 4, the matrix inversion and multiplication consume $O(n^3)$ and $O(n^2d)$ respectively. Suppose the numbers of two types of calculation are p and q, the number of iterations in **Algorithm 1** is r, and the number of subjects is K, the overall computational complexity of this algorithm would be $O(prKn^3) + O(qrKn^2d)$.

Experiments

Baseline: Since our method belongs to embedding model, in this work we make a comparison with six embedding model based algorithms. Specifically, these methods are two unsupervised manifold embedding methods Neighborhood Preserving Embedding (NPE) (He et al. 2005) and Locality Preserving Projections (LPP) (He and Niyogi 2003), and four supervised manifold embedding methods, Supervised Locally Embedded Analysis (SLEA) (Fu and Huang 2006), Supervised Locality Preserving Projections (SLPP) (Li et al. 2007), Supervised Manifold Learning by Chiraz Ben-Abdelkader (SML-B) (BenAbdelkader 2010), and the state-of-the-art work, Supervised Manifold Learning method by Wang *et al*. (SML-W) (Wang and Song 2014). To make a fair comparison, for all the methods, nearest neighbour (NN) is used for pose prediction after learning the embedding coefficients and parameters are fine-tuned to obtain the best performance.

Database: **CMU-PIE** (Sim, Baker, and Bsat 2003) includes 68 subjects of totally 41368 images. Each subject has different poses, expressions and illuminations. As discussed in the previous section, our method takes cubic time complexity in term of sample number in each subject. To reduce the computational cost, the first 15 subjects are used. For each subject, there are 9 different poses in yaw direction varying from -90 degree to 90 degree with the step of 22.5 degree. We crop each image and resize it to the size of 32 $\times$ 32 pixels. **MIT-CBCL** (Rowley, Baluja, and Kanade 1998; Alvira and Rifkin 2001) contains 3D synthetic facial images of 10 subjects. The head models are generated by fitting a morphable model to high-resolution training images. For each subject, there are 9 poses in yaw direction varying from 0 degree to -32 degree at increments of approximate 4 degree. Each pose contains 36 different illuminations. **Extended Yale B** (Georghiades, Belhumeur, and Kriegman 2001) contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. Different from the previous two datasets, these 9 poses are neither in yaw nor pitch direction, without precise pose angles (Please refer to (Georghiades, Belhumeur, and Kriegman 2001) for more details). Instead of precise head pose estimation, we use extended Yale B for classification accuracy evaluation.

Note that in our experiments, we use gray-scale intensity value as input feature, instead of HOG feature as some previous works did (Wang and Song 2014; Haj, Gonzàlez, and Davis 2012). Two major reasons: first, HOG feature calculating the oriented gradients eliminates some noises, which is not preferable in model robustness evaluation. To evaluate the robustness of different methods, six levels of random noise are added, see Figure 2 as an example. Second, from the aspect of computational cost, gray-scale intensity is faster than HOG feature generation. For the data preprocessing, we follow the similar strategy (Liu, Lin, and Yu 2010; Lu et al. 2012) to reduce the data dimension to 200 via PCA for further speedup.

Evaluation Metric: Mean absolute error (**MAE**) is a well-known and popular used metric to evaluate the performance of pose estimation model. It is calculated as MAE $= E[|\mathbf{p} - \mathbf{g}|]$, where $\mathbf{p}$ denotes the predicted pose angles in vector form, and $\mathbf{g}$ denotes the ground-truth angles. The expectation $E[\cdot]$ measures the average error of the predictions. Classification accuracy (**ACC**) is another evaluation metric widely used in head pose estimation field (Haj, Gonzàlez, and Davis 2012; Geng and Xia 2014). It is defined as ACC $= \sum_{i=0}^{N} \delta(\mathbf{p}_i, \mathbf{q}_i)/N$, where $\delta(x, y)$ is the delta function that equals one if $x = y$ and equal zero otherwise, and N is the total number of predicted poses.

MAE measures the average absolute error between prediction and ground-truth. However, an estimation with a small MAE does not mean a high classification accuracy. It happens when the few wrong classifications (i.e., high accuracy) are far from ground-truth, which results in a bad MAE result. With both metrics, a comprehensive evaluation is provided. In our experiment described below, MAE is calculated on CMU-PIE and MIT-CBCL, ACC is computed on CMU-PIE and extended YaleB. As discussed above, ex-

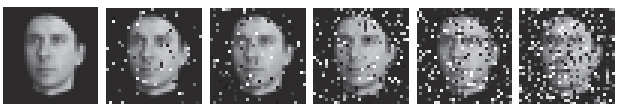


Figure 2: Illustration of noisy training data with different random noise level from original (no noise) to 25% with the step of 5%. Image sample belongs to database MIT-CBCL.

Table 1: MAE (Degree) result on database CMU-PIE.

	0%	5%	10%	15%	20%	25%
NPE	7.451	7.632	8.491	9.670	11.777	12.816
LPP	8.986	9.209	9.335	10.172	11.630	11.926
SLEA	7.472	7.940	8.749	9.105	9.258	10.416
SLPP	7.374	7.584	9.021	9.704	11.281	11.603
SML-B	7.242	7.765	9.014	9.523	9.990	10.695
SML-W	5.547	7.543	7.967	9.159	9.383	10.248
Ours	**4.395**	**4.682**	**5.986**	**6.698**	**7.081**	**7.363**

Table 2: MAE (Degree) result on database MIT-CBCL.

	0%	5%	10%	15%	20%	25%
NPE	0.204	1.737	1.960	2.300	2.541	3.157
LPP	0.107	1.085	1.423	2.515	2.637	3.050
SLEA	0.109	0.873	1.689	2.062	2.076	2.408
SLPP	0.091	1.332	1.780	1.893	1.975	2.427
SML-B	0.080	0.982	1.440	1.678	1.961	2.055
SML-W	0.020	0.442	1.273	1.428	1.766	2.003
Ours	**0.011**	**0.017**	**0.024**	**0.414**	**0.998**	**1.311**

Table 3: Classification ACC (%) on database CMU-PIE.

	0%	5%	10%	15%	20%	25%
NPE	89.64	85.46	83.75	78.05	72.47	72.81
LPP	85.22	83.63	78.67	77.18	74.23	73.77
SLEA	89.36	87.75	86.39	82.39	80.50	78.23
SLPP	87.10	84.93	79.44	76.12	73.86	74.42
SML-B	89.63	87.11	86.74	84.07	83.25	80.43
SML-W	91.77	87.42	88.03	83.91	83.34	80.91
Ours	**91.94**	**89.56**	**90.01**	**86.05**	**83.72**	**82.79**

Table 4: Classification ACC (%) on extended YaleB.

	0%	5%	10%	15%	20%	25%
NPE	79.91	79.58	79.11	78.40	78.21	77.32
LPP	81.01	79.74	79.67	78.17	77.70	77.43
SLEA	81.74	81.50	80.52	79.58	79.90	77.95
SLPP	82.70	81.93	82.21	79.74	80.47	79.41
SML-B	85.08	85.02	84.68	84.11	83.08	82.89
SML-W	86.34	85.96	85.49	85.27	83.65	83.42
Ours	**88.17**	**87.91**	**86.34**	**86.05**	**84.78**	**84.26**

tended YaleB only includes poses but without precise groud-truth angles, so only ACC can be computed. Due to the simplicity of MIT-CBCL, we only show MAE result since ACC is lack of discriminability.

Result

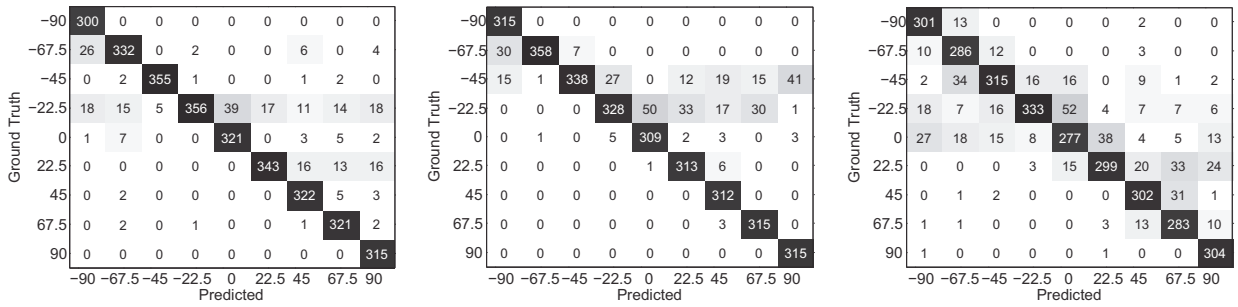
For all experiments, five-fold cross-validation is applied as (Haj, Gonzàlez, and Davis 2012). Table 1 and 2 tabulate the MAE results on databases CMU-PIE and MIT-CBCL with different levels of random noise from 0% (no noise) to 25% at an increment of 5%. According to the observation, several statements can be generalized as follows:

- Our proposed method PLRE consistently outperforms all the other baselines. PLRE reduces the average error across different noise settings from 8.308 degree to 6.034 degree in CMU-PIE and 1.155 degree to 0.463 degree in MIT-CBCL, respectively.
- All the supervised methods SLEA, SLPP, SML-B, SML-W and ours PLRE outperform the unsupervised methods NPE and LPP.
- With more noise, the performances of all methods get worse. However, as the noise level grows, more improvements are achieved by our PLRE, i.e., from 1.152 degree (no noise) to 2.885 degree (25% noise) in CMU-PIE.

Discussion: All the observations are in expectation, as SLEA, SLPP, SML-B, SML-W and ours incorporate supervised information, thus more discriminative embedding coefficients are learned for pose estimation. It is our method PLRE that firstly considers the subject-dominant influence, and formulates pose estimation with different subjects under multi-task learning framework. By uncovering the global structure via low-rank embedding, our method PLRE outperforms all the other competitors.

Table 3 and 4 show the classification accuracy on database CMU-PIE and extended YaleB with different noise levels. Same as MAE, our method PLRE performs best. The similar trend is found that with higher level of noise, all the methods perform worse. However by comparing the reported MAE (Table 1) and ACC (Table 3) on CMU-PIE dataset, we find a case that pose prediction is with better ACC but worse MAE, e.g., SLEA and SLPP under 5% noise condition. It is reported that SLEA and SLPP have 7.940 degree and 7.584 degree respectively, which means SLPP is better than SLEA in term of MAE. While as for ACC, SLEA and SLPP report 87.85% and 84.93%, denoting that SLEA overcomes SLPP. This phenomenon is in accordance with the discussion made in previous subsection. Under both evaluation metrics, PLRE consistently outperforms all the competitors.

To better demonstrate the effectiveness of our method, confusion matrix is used to specify the exact classification result between the estimated pose angle and ground-truth as shown in Figure 3. The matrices are generated on database CMU-PIE with three different noise levels, 0%, 10%, 20%, respectively. Each number represents the frequency of occurrence when prediction is correct. Since five-folder cross-validation is applied, the sum of all numbers in the matrix is the total number of images. Then the accuracy rate can be calculated as the sum of diagonal number divided by the total number, i.e., $\mathrm{ACC} = \sum_i(\mathrm{diag}(M_{ii}))/\sum_{ij}(M_{ij})$, where M denotes the confusion matrix, i and j index the row and column, respectively. As expected, from no noise (Figure 3(a)) to 20% noise (Figure 3(c)), the accuracy drops from 91.94% to 83.72%. We also observe that when small noise is added, our method works very well for large-angle pose estimation. For example, under 0% and 10% noise circumstances, the prediction accuracies are 100% for -90 and 90 degrees. While for more frontal cases, such as -22.5, 0, and 22.5 degrees, there are a lot of misclassifications. This is because side faces are quite different from frontal faces. Moreover, it is difficult to distinguish the near-frontal faces with corrupted face images when large noise is added. Although some exceptions happen, we can still observe the trend from Figure 3(c) that our model produces the close



(a) No Noise (b)10% Noise (c)20% Noise

Figure 3: Confusion Matrices of dataset CMU-PIE with different noise levels, i.e., (a) no noise, (b) 10% and (c) 20%, respectively. The number represents the occurrence frequency between prediction and groundtruth.

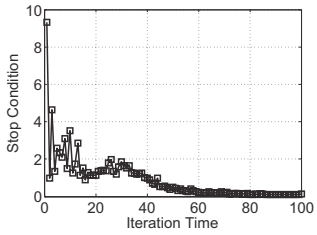


Figure 4: Convergence study of stop condition with respect to iteration time on CMU-PIE database. Here, stop condition is defined as $\max(\|P_{i,t}^{\mathrm{T}}X_i - P_{i,t}^{\mathrm{T}}X_iZ_{i,t} - E_{i,t}\|_\infty, \|Z_{i,t} - J_{i,t}\|_\infty)$.

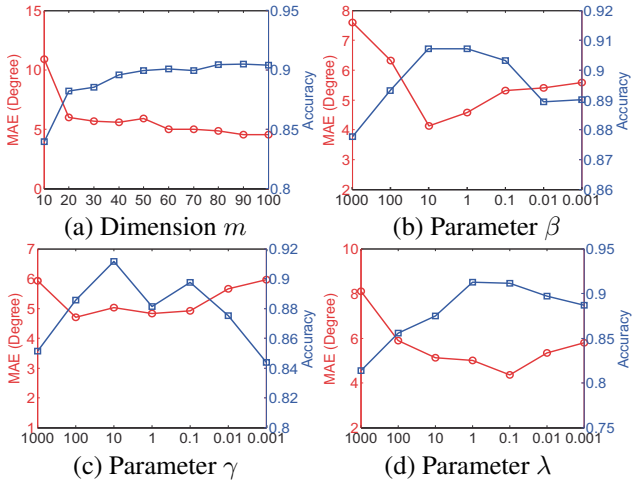
estimations for those misclassification cases.

Convergence and Parameter Analyses

To test the stability and robustness, five experiments are conducted to study the pose estimation performance in terms of convergence and model parameters. Without explicit specification, all the analytical experiments are conducted on CMU-PIE with the parameters set as $m = 100$, $\beta = 10$, $\gamma = 1$ and $\lambda = 0.1$.

Convergence analysis. To show the convergence property, we compute stop condition as $\max(\|P_{i,t}^{\mathrm{T}}X_i - P_{i,t}^{\mathrm{T}}X_iZ_{i,t} - E_{i,t}\|_\infty, \|Z_{i,t} - J_{i,t}\|_\infty)$ in t-iteration, the convergence curve is plotted in Figure 4. From the observation, the whole process can be generalized into two stages. In the first stage (#1 $\sim$ #30), the stop condition fluctuates sharply. Then (after #30), it drops smoothly until the final convergence. Note that with different parameter settings, the convergence curves might not be exactly the same, but the trends are similar. From this experiment, it is well demonstrated that our method is robust from convergence aspect.

Parameter analysis. A coarse-to-fine strategy is adopted to find the proper range of each parameter. In our model, there are four major parameters, i.e., the dimension m of projection matrix, trade-off parameter β for pose-dependent regularizer term, γ for cross-task alignment term and λ for error term. Projection matrix P plays the crucial role in head pose estimation task. Accordingly, analysis on the dimension m of P is critical to show the effectiveness and robustness of our method. We set the dimension m in the range of [10, 100] with a step of 10. Figure 5(a) reports the performance in terms of MAE and ACC results. It is clearly observed that as the dimension m grows, the performance



(a) Dimension m (b) Parameter β

(c) Parameter γ (d) Parameter λ

Figure 5: Parameter analyses on CMU-PIE database. (a) MAE (red curve) and ACC (blue curve) with respect to the dimension m of projection matrix P. (b-d) show the MAE and ACC curves with respect to trade-off parameters β, γ and λ, respectively. The default values of four parameters are set to be $\{100, 10, 1, 0.1\}$.

gets better, and eventually keeps steady when m reaches 90. Note that when m gets larger, the computational cost also becomes more expensive. Thus in our experiments, we set $m = 100$ as default.

Figure 5(b-d) show the performance with respect to three different balancing parameters of PLRE. The values of parameters are selected in the grid of $[10^3, 10^2, 10^1, 1, 10^{-1}, 10^{-2}, 10^{-3}]$. We can observe that when β, γ and λ are set around 10, 1 and 0.1, we get a relatively good MAE and ACC result. Note that there is a drop when $\gamma = 1$ in term of ACC (see Figure 5(c)). However as for MAE, $\gamma = 1$ performs best. This phenomenon is similar with the case of SLEA/SLPP that we discuss above.

Conclusion

In this paper, we proposed a Pose-dependent Low-Rank Embedding (**PLRE**) method for head pose estimation. Inspired by several empirical observations, we designed cross-task alignment term and pose-dependent graph regularizer under low-rank multi-task framework. The superior results on three benchmarks with different levels of random noise had shown our superiority.

Acknowledgments

This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, NPS award N00244-15-1-0041, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

Alvira, M., and Rifkin, R. 2001. An empirical comparison of snow and svms for face detection. A.I. memo 2001-004, Center for Biological and Computational Learning, MIT, Cambridge, MA.

Balasubramanian, V. N.; Ye, J.; and Panchanathan, S. 2007. Biased manifold embedding: A framework for person-independent head pose estimation. In *CVPR*.

BenAbdelkader, C. 2010. Robust head pose estimation using supervised manifold learning. In *ECCV*, 518–531.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM* 58(3):11.

Ding, Z., and Fu, Y. 2014. Low-rank common subspace for multi-view learning. In *ICDM*, 110–119.

Ding, Z.; Shao, M.; and Fu, Y. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, 1192–1198.

Edwards, G. J.; Lanitis, A.; Taylor, C. J.; and Cootes, T. F. 1998. Statistical models of face images - improving specificity. *IVC* 16(3):203–211.

Fanelli, G.; Gall, J.; and Van Gool, L. 2011. Real time head pose estimation with random regression forests. In *CVPR*, 617–624.

Fu, Y., and Huang, T. S. 2006. Graph embedded analysis for head pose estimation. In *IEEE Conference on Automatic Face and Gesture Recognition*, 3–8.

Geng, X., and Xia, Y. 2014. Head pose estimation based on multivariate label distribution. In *CVPR*, 1837–1842.

Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. J. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *TPAMI* 23(6):643–660.

Haj, M. A.; Gonzàlez, J.; and Davis, L. S. 2012. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. In *CVPR*, 2602–2609.

He, X., and Niyogi, P. 2003. Locality preserving projections. In *NIPS*.

He, X.; Cai, D.; Yan, S.; and Zhang, H. 2005. Neighborhood preserving embedding. In *ICCV*, 1208–1213.

He, K.; Sigal, L.; and Sclaroff, S. 2014. Parameterizing object detectors in the continuous pose space. In *ECCV*, 450–465.

Kwong, J. N. S., and Gong, S. 2002. Composite support vector machines for detection of faces across views and pose estimation. *IVC* 20(5-6):359–368.

Li, S., and Fu, Y. 2014. Robust subspace discovery through supervised low-rank constraints. In *SDM*, 163–171.

Li, Z.; Fu, Y.; Yuan, J.; Huang, T. S.; and Wu, Y. 2007. Query driven localized linear discriminant models for head pose estimation. In *ICME*, 1810–1813.

Lin, Z.; Chen, M.; and Ma, Y. 2009. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. In *Technical Report, UIUC*.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. *TPAMI* 35(1):171–184.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *ICML*, 663–670.

Lu, C.-Y.; Min, H.; Zhao, Z.-Q.; Zhu, L.; Huang, D.-S.; and Yan, S. 2012. Robust and efficient subspace segmentation via least squares regression. In *ECCV*, 347–360.

Murphy-Chutorian, E., and Trivedi, M. M. 2009. Head pose estimation in computer vision: A survey. *TPAMI* 31(4):607–626.

Rowley, H. A.; Baluja, S.; and Kanade, T. 1998. Neural network-based face detection. *TPAMI* 20(1):23–38.

Shao, M.; Kit, D.; and Fu, Y. 2014. Generalized transfer subspace learning through low-rank constraint. *IJCV* 109(1-2):74–93.

Sim, T.; Baker, S.; and Bsat, M. 2003. The cmu pose, illumination, and expression database. *TPAMI* 25(12):1615–1618.

Wang, S., and Fu, Y. 2015. Locality-constrained discriminative learning and coding. In *CVPR Workshops*, 17–24.

Wang, C., and Song, X. 2014. Robust head pose estimation via supervised manifold learning. *Neural Networks* 53:15–25.

Wang, J., and Sung, E. 2007. EM enhancement of 3d head pose estimated by point at infinity. *IVC* 25(12):1864–1874.

Wang, X.; Huang, X.; Gao, J.; and Yang, R. 2008. Illumination and person-insensitive head pose estimation using distance metric learning. In *ECCV*, 624–637.

Wang, Y.; Xu, H.; and Leng, C. 2013. Provable subspace clustering: When LRR meets SSC. In *NIPS*, 64–72.

Yang, J.; Yin, W.; Zhang, Y.; and Wang, Y. 2009. A fast algorithm for edge-preserving variational multichannel image restoration. *SIAM Journal on Imaging Sciences* 2(2):569–592.

Zhao, H., and Fu, Y. 2015. Dual-regularized multi-view outlier detection. In *AAAI*, 4077–4083.