Two Dimensional Partitioned Sparse Representation for Head Pose Estimation

Chao ZHANG*, Yanning ZHANG†, and Liang LIAO‡
*†‡Shaanxi Provincial Key Laboratory of Speech and Image Information Processing
Northwestern Polytechnical University, Xi'an
‡School of Electronic and Information Engineering
Zhongyuan University of Technology, Zhengzhou

***Abstract*—Although classical sparse representation is capable to solve appearance based classification problems such as face recognition, it is problematic that images need to be converted to column vectors before subsequent processing which makes the computation expensive due to the huge dimension. From the human vision perspective, it is reasonable to observe image in form of matrix rather than vector. To reduce the computational complexity, the idea to partition the images is introduced as well. We combine partition processing with two dimensional sparse representation together to propose 2DPSRC (2D Partitioned Sparse Representation Classifier) considering the property of head pose estimation problem. It can greatly improve the estimation accuracy and enhance the efficiency of the computational process involved pursuit of ℓ_1-norm minimization. Finally, experiments on Pointing'04 and Oriental Face Database show the effectiveness and robustness of our proposed method.**

I. Introduction

Head pose estimation is a natural step for bringing the information gap between people and computers. This fundamental human ability provides rich information about the intent and motivation which are useful in many real applications, such as multi-view face recognition, human computer interaction, and human-centered scene understanding [1]. It is still challenging to estimate the head pose automatically and robustly using still images. A common approach of head pose estimation is to learn the head pose from a set of face images with the known yaw and pitch angle as class labels. As a result, the unknown pose of test face image can be obtained by applied the learned classifiers.

In head pose estimation, the training data can be written in forms of $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_l, \mathbf{y}_l)\}, \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^2$ where $\mathbf{x}_i$ is the representation of a face image, and $\mathbf{y}_i$is a pose label, either the horizontal angle or the vertical angle. If each pose label is considered to be a single class, the head pose estimation becomes a classification problem. On the other hand, once the possible pose angles are ordered, this problem can also be thought of as a classical regression problem.

In this paper, we consider the head pose estimation problem as a classification problem rather than a regression one. It is simple and suitable to apply SRC to tackle this problem. The motivation of this study is our previous work on face recognition [5]. We found that SRC is capable to solve appearance based head pose classification problem as well. A promising method is to extend the vector-based sparse representation to matrix-based sparse representation approach for the better performance in face recognition [9]. Here, we want to investigate whether the same result could be obtained from another problem - head pose estimation. We compare the classical SRC and proposed 2DPSRC in both yaw and pitch estimation experiments. To evaluate the proposed head pose estimation method, we conduct evaluation experiments thoroughly on the Pointing'04 head pose database [7]. This public available database contains rich numbers of head poses, e.g., (-90, -75, -60, -45, -30, -15, 0, +15, +30, +45, +60, +75, +90) in horizontal direction, and (-90, -60, -30, -15, 0, +15, +30, +60, +90) in vertical direction. Furthermore, we also make use of Oriental Face Database to exploit the characteristics of 2DPSRC using different selection of partition number and dimension.

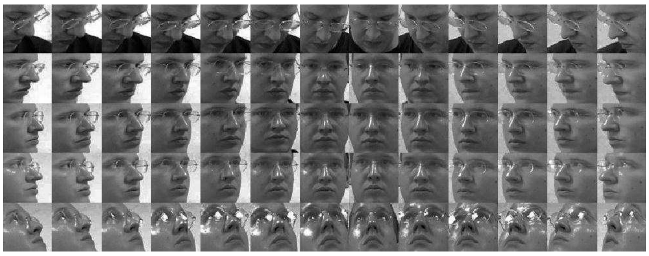

Fig. 1. Samples of multi-pose face images in Pointing '04 Database. From top to bottom, pitch varies from -60 to 60 degrees. From left to right, yaw changes from -90 to 90 degrees

In the remaining of the paper, we first briefly review the sparse representation and indicate how it is suitable to head pose estimation. Then our methodology was presented with analysis in Section III. Experiments results are described in Section IV.

II. Related Work

After the introduction of compressive sensing in image processing and pattern recognition [2], SRC was first proposed by [4] by Wright et al. Although quite easy to understand, SRC reveals the essential principle that if the training samples are sufficient to represent the variation of all possible factors, any test sample could be written as the linear combination of those training samples related to the same class. Another popular idea is to utilize patch-based method to solve appearance based pattern classification problem such as face recognition and pose estimation [10][11]. In real world application, even the resolution of an single image is very huge, a under sampling image of it is also sufficient for human vision to capture the primary information with little loss of detail information. Seeing the successful application of SRC in face recognition, we intend to apply this promising method in head pose estimation problem which has great similarity as face recognition.

Considering a training data set $\{(\mathbf{x}_i, \mathbf{y}_i); i = 1, \ldots, n; \mathbf{x}_i \in \mathbb{R}^d, \mathbf{y}_i \in \{1, 2, \ldots, N\}\}$, where $\mathbf{x}_i$ represents the ith sample, a d-dimension column vector, and$\mathbf{y}_i$ is the label of the ith sample with N as the number of classes. For a testing sample $\mathbf{y} \in \mathbb{R}^d$, the problem of sparse representation is to find a column vector $\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$ such that $\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_n\mathbf{x}_n$ while at the same time $\|\mathbf{a}\|_0$ is minimized, where $\|\mathbf{a}\|_0$ is $l_0 - norm$, and it is equivalent to the number of nonzero components in the vector $\mathbf{a}$.

Defining a matrix by putting $\mathbf{x}_i$ as the ith column $\mathbf{X} = [x_1, x_2, \ldots, x_n]$, the problem of sparse representation can be converted to:

$$\mathbf{a} = \min \|\mathbf{a}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{X}\mathbf{a}. \tag{1}$$

Since the solution to sparse representation problem is NP-hard due to its nature of combinational optimization. Also, ℓ_1-norm minimization can efficiently recover the sparse signal and is robust against outliers. Therefore, the problem is reduced to solve an ℓ_1-norm minimization problem. Here, we use the L1 Magic [7] software package to get the optimum solution.

In our previous work, we also extend the work to 3D face recognition by generating range image from 3D model using a novel grid down sampling method. Similar to the face recognition problem, head pose estimation problem is considered to be a classification problem which can be handled by ℓ_1-norm minimization. [4] validated the assumption that head pose estimation can obtain satisfactory result under the sparse representation framework, also, they proposed BSRC to reduce the influence of background information.

There is still limitation in estimating head pose solely using SRC. As under the traditional SRC framework, each image is stacked as a column vector, undoubtedly leading to the loss of structure information which is essentially related to poses information. Moreover, once the size of the image is too huge, e.g. 1024×768, it is impossible to pursue the possible solution using PC. Thus, the dimension reduction technique is necessary to alienate the computational complexity such as down sample or PCA method. However, whether dimension reduction will affect or lower the accuracy is still an open problem.

Instead of relying on dimension reduction, we propose to extend vector-based SRC to matrix-based SRC in order to avoid the curse of dimensionality, and exploit the neighboring information of pixels in head pose estimation. Another difference from traditional SRC is that we can assign a pose label to the test sample according to the diagonal elements rather than the residuals between the original test image and the reconstructed image. This difference can render the proposed method more robust to misalignment error and more flexible in afterward processing.

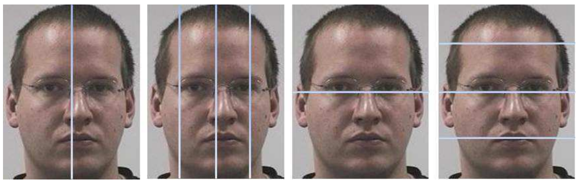

Fig. 2. Illustration of the partition method.
Due to the limitation of pages, selected samples only present the variation when the partition number is chose to be 2 or 4 in both horizon and vertical direction.

III. 2D Partitioned SRC for Head Pose Estimation

A. Motivation and Analysis

Traditional classification based on sparse representation rely on the vector form of image. Given sufficient training samples of the ith class, $\mathbf{A} = [\mathbf{v}_{i,1}, \mathbf{v}_{i,2}, \ldots, \mathbf{v}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$, any new sample $\mathbf{y} \in \mathbb{R}^m$ from the same class will approximately lie in the linear span of the training samples associated with object i: $\mathbf{y} = \alpha_{i,1}\mathbf{v}_{i,1} + \alpha_{i,2}\mathbf{v}_{i,2} + \ldots + \alpha_{i,n_i}\mathbf{v}_{i,n_i}$. The linear representation of $\mathbf{y}$ can be written in terms of all training samples as:

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 \in \mathbb{R}^m. \tag{2}$$

where $\mathbf{x}_0 = [0, 0, 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,n_i}, 0, \ldots, 0] \in \mathbb{R}^n$ is a coefficient vector whose entities are zero except those who are associated with the ith class. The ratio of sparseness is $1/K$ where K is the number of classes.

However, in our proposed 2DPSRC framework, each column of the new sample $\mathbf{Y} \in \mathbb{R}^{h \times w}$ will lie in the linear span of the training samples associated with the object i at the jth column, the test sample could be written in following form:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \in \mathbb{R}^{w \times h}, \tag{3}$$

with the coefficient matrix as:

$$\mathbf{X} = [0, 0, \alpha_{i,1,j}, 0, 0, \alpha_{i,2,j}, 0, 0, \alpha_{i,3,j}, 0, 0]^T. \tag{4}$$

Thus, the ratio of sparseness becomes

$$\begin{aligned} \lambda &= (w \times n)/(w^2 \times n \times K) \\ &= 1/(K \times w) \end{aligned} \tag{5}$$

We notice that the number of columns affects the degree of sparseness. Based on the assumption of sparse representation, our proposed method can effectively improve the classification accuracy. Even though, adopting all the columns tends to suffer from great inefficiency. In order to strike a balance between the efficiency and effectiveness, we choose to divide the image into divisions e.g. 2 or 4 parts. Our experiments results could support our analysis. Besides, the underdetermination characteristic plays an key role in the solution of the optimum

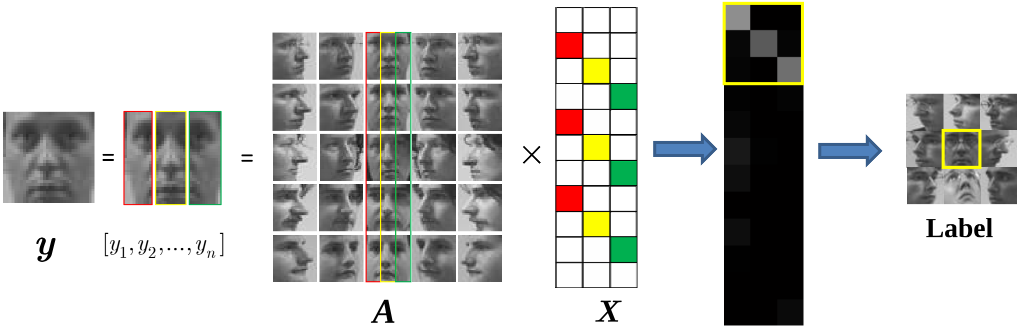

Fig. 3. Flow chart of the algorithm.

solution in ℓ_1-norm minimization, and then the dimension of image must be less than the number of training samples. It is also known that as the dimension increases, appearance based classification is expected to obtain higher accuracy. However, small sample problem is common due to the obstacle to collect sufficient data. To apply 2DPSRC can effectively overcome this weakness. Here, we define the number of training sample as N, the dimension of image is D. If we choose to divide the image into p parts. We denote ratio ρ as follows:

$$\rho = (N \times p) \div \frac{D}{p} = \frac{N}{D} \times p^2 \tag{6}$$

Thus, the introduction of ρ can improve the underdetermination of the dictionary greatly.

B. Algorithm Description

Based on the above analysis, 2DPSRC is proposed in this paper. Fig. 1 shows the flow chart of the method. For the given face images $\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_p$ are generated by partitioning the image into p parts. In yaw angle estimation, the image is divided horizontally, while in pitch estimation, vertically dividing is applied.

In the second step, for each part of the image $\mathbf{I}$, the features $\mathbf{x}_i$ can be obtained by using dimension reduction methods to get the compact representation of the original images data.

In the third step of our method, each part of training or test images is converted to a column vector, and then SRC is applied to get the sparse coefficient matrix $\mathbf{X}$. The matrix $\mathbf{X}$ is composed of K sub matrixes which represent the distribution of the given image in each class. There are two means to get the final pose label: minimizing the residual score or maximizing the diagonal summation as the elements in the diagonal line should be large nonzero values. In the former manner, similar to the SRC, the residual γ is computed and the label is given by choosing the class corresponding to the minimum one. In the latter one, we summate diagonal elements of each sub matrixes to arrive a row vector $[v_1, v_2, ..., v_K]$, the class label is predicted by maximizing the summation of the class-dependent diagonal elements:

$$L = L_i, \quad s.t. \quad i = \max v_i, i = 1, 2, \ldots, K. \tag{7}$$

or

$$L = L_i, \quad s.t. \quad i = \min \gamma_i, i = 1, 2, \ldots, K. \tag{8}$$

TABLE I

Algorithm 1
(2D Partitioned Sparse Representation Classification).

INPUT:
The test face images $\mathbf{I}$. The size of $\mathbf{I}$ is $h \times w$, where h is the height and w is the width of the image;
STEP 1:
Divide the images into N parts $\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_n$;
STEP 2:
For the part $\mathbf{I}_i$, getting its feature $\mathbf{x}_i$. Then, each image is represented using a series feature $\mathbf{x}_i$;
STEP 3:
For the feature $\mathbf{x}_i$, getting its sparse representation $\mathbf{y}_i$,and then, the sparse matrix is obtained by stacking $\mathbf{y}_i$;
STEP 4:
Compute the sparse matrix, and arrive the summation λ of diagonal values or the residual γ of the reconstruction.
Get the predicted pose label l by maximizing the summation or minimizing the residual;
OUTPUT:
The predicted pose label l.

IV. Experiments and Analysis

A. Database

In order to validate the effectiveness of our proposed method, we conduct experiments on two public face databases. The first one is the Pointing '04 database [6]. The head pose database consists of 15 sets of images. Each set contains of 2 series of 93 images of the same person at different poses. Naturally, the first part of the database can be used for training purpose, while the second part could be used as test database. The pose is determined by 2 angles (h, v), which varies from -90 degrees to +90 degrees in yaw and from -60 degrees to + 60 degrees in pitch. The images are cropped into the same size of 64×64.

Another database we use is The Oriental Face Database [8] which contains 33669 face images of 1247 individuals. These images are divided into two databases: the viewpoint face database and the illumination face database. In our experiment, we use a subset of the database including data of 78 individuals, ranging from -90 degrees to +90 degrees (images are taken every 10 degrees).

B. *Experiment on Pointing ’04 Database*

As the Pointing ’04 Database is rich in poses both in yaws and in pitches, we design both experiments to test the effectiveness of our proposed method. Moreover, comparisons between traditional SRC method and 2DPSRC are given to clearly show the advantageous performance.

In the experiments of yaw estimation, we group the images according to their pose labels. More specifically, they are divided into 13 groups (each group contains 105 images for training or testing), from -90 degrees to 90 degrees. We random choose a subset from the training set and also conduct the testing in the same manner. For example, for each pose, we random choose 7 images from each individual to build the dictionary (the dictionary is composed of 7×15×13=1365 images), and then the rest images (6×15×13=1170) in the test group are used to estimate poses. Hence, there is no overlapping in the training and testing images we use.

Similarly, in the experiments of pitch estimation, both training set and testing set are divided into 7 classes each of which represents a pose belonging to the set of $-60, -30, -15, 0, +15, +30, and + 60$. The organization of the data is the same as in yaws experiments.

For both yaw and pitch experiments, we conduct experiments using various ratio of training number and testing number, e.g. 7 for training vs. 6 for testing, 6 for training vs. 7 for testing, 5 for training vs. 8 for testing in yaw experiments. Then, the average MAE is obtained by averaging those different experiments results.

According to Table II, we can see that in the experiments of yaw estimation, results of 2DPSRC using either residual minimization or diagonal maximization criteria is better than that of 1DSRC. Generally, 2DPSRC method combined with random dimensionality reduction method obtain lower MAE on average over various settings of dimensionality and partition number. To be more specifical, the best result for this experiments is obtained using 2DPSRC plus the diagonal values maximization using PCA, at the MAE of 11.72.

From the pitch experiments results shown in Table III, it is not surprising to observe that performance of 2DPSRC exceeds the counterpart of the traditional 1DSRC. More exactly, when the dimension reduction is conducted by using either random sampling or PCA, the estimation MAE is lower than 1DSRC, while due to the loss of discriminative information in the process of simple down sampling, 2DPSRC did not express the strength to beat 1DSRC. Moreover, it seems that PCA obtain more satisfactory results among all three kinds of dimension reduction methods, the reason lies in that the feature extraction is done through a supervised way. Another finding is that the performance of 2DPSRC using different classification criteria is similar which reveals that the sparse coefficient is strong enough to be classified correctly. PCA is considered to be a more suitable dimensionality reduction method compared to others. The best result is obtained by 2DPSRC plus diagonal using PCA, at the MAE of 9.76.

C. *Experiment on Oriental Face Database*

In order to confirm the effectiveness and exploit the strength of the proposed method, we utilize the Oriental Face Database to evaluate the results which contains more identities and thus is more challenging. The Oriental Face Database contains 78 persons, with each person including 19 poses in horizontal variations. In the experiment, we random choose 39 persons as training set, and another 39 persons for testing purpose. We compare the results using three kinds of dimension reduction method, as well as four dimensions. 2 parts and 4 parts are both used in the experiments.

Considering the computational concerns and huge number of training images and testing images, we choose to compare two selections of partition: 2 and 4. As to the dimension, we select the down sample rate to ensure that the dictionary can meet the under determination requirement. Hence, 1024, 512, 256, and 128 are evaluated according to the down sample rate of 1/4, 1/8, 1/16 and 1/32. The results show that the 4 parts beat 2 parts in most cases, which proves our prediction. Again, random sampling and PCA are better than down sampling strategy.

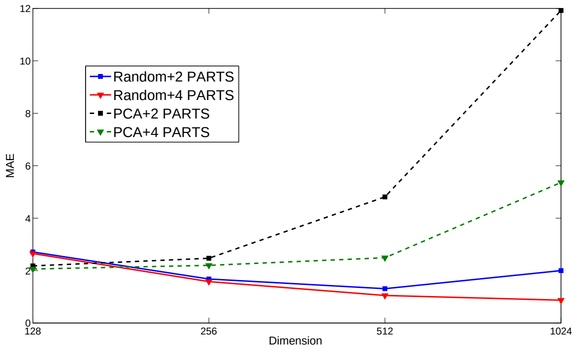

Fig. 4. Evaluation of dimension and partition number on Oriental Face Database.

The most remarkable of the proposed method is that as the dimension increases, it is not guaranteed that the performance can also improve, as we think, this is due to the constrainment that the assumption of underdetermination of the dictionary. Because of the angle is each class is 10 degrees, the MAE is much lower that the experiments on Pointing ’04 in which the degree is 15. Another reason is maybe that images in the Oriental Face Database are without variation in vertical angles. Among these experiments, the best result is obtained using random sampling method at the dimension of 1024.

V. Conclusion

In this paper, we validate that the classical vector-based SRC could be effectively extended to matrix-based SRC combined with the partitioning method. Consequently, it could preserve

TABLE II
Quantitative evaluation of yaw estimation on Pointing '04 database.

Method	1DSRC			2DPSRC+ResMin			2DPSRC+DiagMax		
Dimension / Partitions	Down	Random	PCA	Down	Random	PCA	Down	Random	PCA
256 / 2	26.04	21.38	22.95	25.00	14.61	18.75	26.98	13.61	13.07
256 / 4	24.06	24.21	24.30	19.63	12.12	14.55	31.09	12.33	11.72
128 / 2	25.09	17.96	13.69	26.20	13.16	13.29	29.03	12.81	12.12
128 / 4	25.15	19.06	14.36	26.58	12.67	13.61	38.15	12.33	12.54

TABLE III
Quantitative evaluation of pitch estimation on Pointing '04 database.

Method	1DSRC			2DPSRC+ResMin			2DPSRC+DiagMax		
Dimension / Partitions	Down [1]	Random [2]	PCA	Down	Random	PCA	Down	Random	PCA
256 / 2	18.95	17.02	12.88	16.61	11.47	12.23	19.23	11.21	10.47
256 / 4	17.40	15.90	13.04	18.61	11.14	10.85	21.11	11.52	10.07
128 / 2	21.59	13.42	10.21	25.23	10.83	10.85	27.61	10.71	9.76
128 / 4	21.26	13.12	11.28	28.45	11.57	10.12	33.54	11.40	10.64

[1] Down Sampling
[2] Random Sampling

the column or row structure information which is related to pose variations, and thus achieving better MAE on various experiments. The introduction of partition method not only improve the estimation accuracy, but also make the SRC based method more applicable when the dictionary does not hold the underdetermination matrix. Dividing the images horizontally or vertically can cope with head pose estimation in yaw or pitch separately.

Due to the sparsity become more remarkable if represented in terms of 2DPSRC, the head pose estimation accuracy can be greatly improved, even using only the information embedded in the diagonal elements. In addition, we exploit the optimum selection of the parameters including partition number and dimensionality. Experiments results show that the method outperforms traditional 1DSRC in both yaw estimation and pitch estimation.

VI. Limitations and Future Work

As mentioned earlier, a significant drawback of 2DPSRC is that the assumption of the sparse representation cannot always hold. Actually, the data collected in real application could not be as sufficient as data in face database. Then, the sparse coefficient we get is dispersed instead of concentrated on the related class using the L1-norm minimization. We plan to use alternative method to overcome this problem.

There are some future work we can do to improve our proposed method. First, we reckon 2DPSRC could probably be further exploited by applying two directional compressive sampling proposed in [9]. It is reported that 2DCS can future reduce the computational complexity both in the compressing and in the reconstruction procedure. Second, to replace original image data with more distinctive feature is expected to show better performance in the pose estimation problem, such as SIFT descriptor and LGBP. We consider the situation worth of future investigation.

Acknowledgment

This work is supported by the National High Technology Research and Development Program (863) of China (No.2009AA01Z315), National Natural Science Foundation of China (No.60872145), and Cultivation Fund of the Key Scientific and Technical Innovation Project, Ministry of Education of China (No.708085).

References

[1] E. Murphy-Chutorian, M. M. Trivedi, "Head pose estimation in computer vision: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.31, no.4, pp.607-626, April 2009

[2] E. Candès, M. B. Wakin, "An introduction to compressive sampling," *Signal Processing Magazine, IEEE* , vol.25, no.2, pp.21-30, March 2008

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Yi Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.31, no.2, pp.210-227, Feb. 2009

[4] Bingpeng Ma, Tianjiang Wang , "Head pose estimation using sparse representation," *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on* , vol.2, no., pp.389-392, 19-21 March 2010

[5] Chao Zhang, Yanning Zhang, Zenggang Lin, Zhe Guo, "An efficiently 3D face recognizing method using range image and sparse representation," *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on* , vol., no., pp.1-4, 10-12 Dec. 2010

[6] N. Gourier, D. Hall, J. L. Crowley , "Estimating face orientation from robust detection of salient facial features," *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK

[7] Candès, E., "L1-Magic: recovery of sparse signals," *http://www.acm.caltech.edu/l1magic/*

[8] Oriental Face Database, *http://www.aiar.xjtu.edu.cn/groups/face/Chinese /Homepage.htm*

[9] Liang Liao, Yanning Zhang, Chao Zhang, 2DCS: Two dimensional random underdetermined projection for image representation and classification, Accepted, *to appear in the proceedings of the 2nd international multimedia technology conference (ICMT2011)*, July, 2011, Hangzhou, China

[10] J. Aghajanian and S. J. D. Prince, "Face pose estimation in uncontrolled environments", *BMVC*, 2009

[11] Zihan Zhou, A. Ganesh, J. Wright, Shen-Fu Tsai, Yi Ma, "Nearest-Subspace Patch Matching for face recognition under varying pose and illumination," *Automatic Face and Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on* , vol., no., pp.1-8, 17-19 Sept. 2008