Research on Image Feature Extraction Method Based on Orthogonal Projection Transformation of Multi-task Learning Technology

Xiaoyuan Jing[1,2], Li Li[1]*, Cailing Wang[1],Yongfang Yao[1],Fengnan Yu[1]
[1]College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China
[2]State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China
*Corresponding author: sagiv_lee@126.com

***Abstract*—When the number of labeled training samples is very small, the sample information we can use would be very little. Because of this, the recognition rates of some traditional image recognition methods are not satisfactory. In order to use some related information that always exist in other databases, which is helpful to feature extraction and can improve the recognition rates, we apply multi-task learning to feature extraction of images. Our researches are based on transferring the projection transformation. Our experiments results on the public AR, FERET and CAS-PEAL databases demonstrate that the proposed approaches are more effective than the general related feature extraction methods in classification performance.**

Keywords-multi-task learning; projection transformation; feature extraction; face recognition

I. INTRODUCTION

The performance of single task for classification cannot meet our requirement when the number of labeled training samples is very small. When there are relations between the tasks to learn, it can be advantageous to learn all tasks simultaneously instead of following the more traditional approach of learning each task independently of the others. There has been a lot of experimental work showing the benefits of such multi-task learning relative to individual task learning when tasks are related, see [1, 2]. Multi-task learning is a very interesting field, which has been used in many fields of pattern recognition in recent years. With multi-task learning, we can utilize the correlation information of several related tasks to learn a few tasks. By this way, we can improve the efficiency of learning tasks, and prompt the recognition rate under condition of SSS issues [3,4,5].Multi-task learning was proposed by Caruana in 1997, a multi-task learning mechanism [6, 7], which train a single multi-layer perceptron to perform multi-task. Ghosn and Bengio [8] proposed a idea based on manifold learning. On the other hand, Silver etc [9, 10] also puts forward a kind of multi-task study, which is given continuously. The challenge of this method probably appears interference of disaster or be forgotten. Argyriou [11] etc proposed a technique for dimension reduction, trying to find a representation in all the tasks of the low dimensional feature space. Jebara [12] proposed an idea, which utilizes the choice of kernel function and feature with support vector machine (SVM) to solve the problem of multi-task learning. Obozinski [13] etc researched the same problem with Argyriou in the view of feature selection.

We propose a new approach to extract features from images with multi-task based on orthogonal projection transformation, which extracts orthogonal feature vectors with related information among several datasets.

II. ORTHOGONAL PROJECTION TRANSFORMATION OF MULTI-TASK LEARNING TECHNOLOGY

A. Supervised feature extraction method based on orthogonal projection transformation

Assume that there are three data set A, B, C, and they are related.Based on the Fisher criterion,we firstly calculate the discriminant transformation W_A on set A through (1)

$$\max_{W_A} J(W_A) = \frac{\left|W_A^T S_{bA} W_A\right|}{\left|W_A^T S_{wA} W_A\right|}. \tag{1}$$

That is to say, we get W_A through (2)

$$P_A W_A = \lambda W_A \,,\; P_A = S_{wA}^{-1} S_{bA}\,. \tag{2}$$

In order to eliminate data the correlation of discriminant transformation between dataset A and B .We set $W_B^T W_A = 0$

$$\max_{W_B} J(W_B) = \frac{\left|W_B^T S_{bB} W_B\right|}{\left|W_B^T S_{wB} W_B\right|}, \tag{3}$$

$$s.t.\; W_B^T W_A = 0$$

According to a general theorem [14], we establish the following objective function:

Theorem 1:

$$\max_{W_2} J(W_2) = \frac{\left|W_2^T S_b W_2\right|}{\left|W_2^T S_w W_2\right|}, \tag{4}$$

$$s.t.\; W_2^T W_1 = 0$$

We get W_A through (5)

$$\begin{aligned} &PW_2 = \lambda W_2, \\ &P = S_w^{-1}\left(I - W_1\left(W_1^T S_w^{-1} W_1\right)^{-1} W_1^T S_w^{-1}\right) S_b, \end{aligned} \tag{5}$$

Where I is a unit matrix, and W_2 is a matrix that consists of d eigenvectors corresponding to d different nonzero eigenvalues of P .

Proof: we construct following Lagrange function

$$L(W_2) = W_2^T S_b W_2 - \lambda\left(W_2^T S_w W_2 - C_1\right) - \mu\left(W_2^T W_1 - C_2\right), \tag{6}$$

where λ and μ are the Lagrange multipliers, and C_1 and C_2 are two constant matrices.

We set the derivative of $L(W_2)$ in (6) on W_2 to be zero:

$$\frac{\partial L(W_2)}{\partial W_2} = 2S_bW_2 - 2\lambda S_wW_2 - \mu W_1 = 0. \tag{7}$$

Multiplying (7) by $W_1^T S_w^{-1}$, we have

$$2W_1^T S_w^{-1} S_b W_2 - \mu W_1^T S_w^{-1} W_1 = 0. \tag{8}$$

Thus μ may be expressed as

$$\mu = 2\left(W_1^T S_w^{-1} W_1\right)^{-1} W_1^T S_w^{-1} S_b W_2. \tag{9}$$

Due to (7) and (9), we have

$$S_bW_2 - \lambda S_wW_2 - W_1\left(W_1^T S_w^{-1} W_1\right)^{-1} W_1^T S_w^{-1} S_b W_2 = 0, \tag{10}$$

That is to say

$$S_w^{-1}\left(I - W_1\left(W_1^T S_w^{-1} W_1\right)^{-1} W_1^T S_w^{-1}\right) S_b W_2 = \lambda W_2, \tag{11}$$

Similarly, we get W_B on dataset B through (12)

$$\begin{aligned} &P_B W_B = \lambda W_B, \\ &P_B = S_{wB}^{-1}\left(I - W_A\left(W_A^T S_{wB}^{-1} W_A\right)^{-1} W_A^T S_{wB}^{-1}\right) S_{bB}, \end{aligned} \tag{12}$$

In order to eliminate correlation of discriminant transform between data set C and data set A, and between data set C and data set B. we construct the objective function and constraint:

$$\begin{aligned} &\max_{W_C} \; J(W_C) = \frac{\left|W_C^T S_{bC} W_C\right|}{\left|W_C^T S_{wC} W_C\right|} \;, \\ &s.t. \quad W_C^T W_A = 0, W_C^T W_B = 0 \end{aligned} \tag{13}$$

$W = [W_A, W_B], W_C, S_{bC}, S_{wC}$ replace W_1, W_2, S_b, S_w respectively, (4) convert into(13).

Similarly ,we get W_C through (14):

$$\begin{aligned} &P_C W_C = \lambda W_C, \\ &P_C = S_{wC}^{-1}\left(I - W\left(W^T S_{wC}^{-1} W\right)^{-1} W^T S_{wC}^{-1}\right) S_{bC}, \end{aligned} \tag{14}$$

Where $W = [W_A, W_B]$.

B. *Unsupervised feature extraction method based on orthogonal projection transformation*

Based on the Local projection preserve criterion, we firstly calculate the discriminant transformation W_A in set A through (15).

$$\begin{aligned} &\min_w \; w^T X_A L X_A^T w \\ &s.t. \; w^T X_A D X_A^T w = 1 \end{aligned} \tag{15}$$

Where $X_A = [x_{A1}, x_{A2}, \cdots, x_{AN}]$ is training samples of data set A, D is a diagonal matrix, $D_{ii} = \sum_j S_{ji}$, S is a similarity matrix, $L = D - S$ .Thus, we get discriminant transformation W_A through (16)

$$X_A L_A X_A^T W_A = \lambda X_A D_A X_A^T W_A \tag{16}$$

Where W_A is a matrix that consists of d_A eigenvectors corresponding to d_A different nonzero eigenvalues of $(X_A D_A X_A^T)^{-1} X_A L_A X_A^T$.

In order to eliminate data the correlation of discriminant transform between set A and B data set. We set $W_B^T W_A = 0$, where W_B is discriminant transform through the Local projection preserve criterion in data set A,so we construct following objective function and constraint:

$$\begin{aligned} &\max_{W_B} \; J(W_B) = \frac{\left|W_B^T X_B L_B X_B^T W_B\right|}{\left|W_B^T X_B D_B X_B^T W_B\right|}, \\ &s.t. \; W_B^T W_A = 0 \end{aligned} \tag{17}$$

Where $X_B = [x_{B1}, x_{B2}, \cdots, x_{BN}]$ is training samples of data set B, D_B is a diagonal matrix, $D_{ii} = \sum_j S_{ji}$, S is a similarity matrix, $L = D - S$,it is a Laplacian Matrix.

Theorem 2:

$$\begin{aligned} &\max_{W_2} \; J(W_2) = \frac{\left|W_2^T X_2 L_2 X_2^T W_2\right|}{\left|W_2^T X_2 D_2 X_2^T W_2\right|}, \\ &s.t. \; W_2^T W_1 = 0 \end{aligned} \tag{18}$$

Where W_1 , X_2 , L_2 and D_2 are known matrices. We get W_A through (19)

$$\begin{aligned} &P W_2 = \lambda W_2, \\ &P = (X_2 D_2 X_2^T)^{-1} \\ &*\left(I - W_1\left(W_1^T (X_2 D_2 X_2^T)^{-1} W_1\right)^{-1} W_1^T (X_2 D_2 X_2^T)^{-1}\right) \\ &*X_2 L_2 X_2^T, \end{aligned} \tag{19}$$

Where I is a unit matrix. W_2 is a matrix that consists of d eigenvectors corresponding to d different nonzero eigenvalues of P .

Proof: we construct following Lagrange function:

$$\begin{aligned} &L(W_2) = W_2^T X_2 L_2 X_2^T W_2 - \\ &\lambda\left(W_2^T X_2 D_2 X_2^T W_2 - C_1\right) - \mu\left(W_2^T W_1 - C_2\right), \end{aligned} \tag{20}$$

where λ and μ are the Lagrange multipliers, and C_1 and C_2 are two constant matrices.

We set the derivative of $L(W_2)$ in (20) on W_2 to be zero:

$$\begin{aligned} &\frac{\partial L(W_2)}{\partial W_2} = 2X_2 L_2 X_2^T W_2 - \\ &2\lambda X_2 D_2 X_2^T W_2 - \mu W_1 = 0. \end{aligned} \tag{21}$$

Multiplying (21) by $W_1^T (X_2 D_2 X_2^T)^{-1}$, we have

$$\begin{aligned} &2W_1^T \left(X_2 D_2 X_2^T\right)^{-1} X_2 L_2 X_2^T W_2 - \\ &\mu W_1^T \left(X_2 D_2 X_2^T\right)^{-1} W_1 = 0. \end{aligned} \tag{22}$$

Thus μ may be expressed as

$$\mu = 2\left(W_1^T\left(X_2D_2X_2^T\right)^{-1}W_1\right)^{-1}W_1^T\left(X_2D_2X_2^T\right)^{-1}X_2L_2X_2^TW_2. \quad (23)$$

Due to (21) and (23), we have

$$\begin{aligned}&X_2L_2X_2^TW_2 - \lambda X_2D_2X_2^TW_2 - \\ &W_1\left(W_1^T\left(X_2D_2X_2^T\right)^{-1}W_1\right)^{-1}W_1^T\left(X_2D_2X_2^T\right)^{-1}X_2L_2X_2^TW_2 = 0,\end{aligned} \quad (24)$$

That is to say

$$\begin{aligned}&(X_2D_2X_2^T)^{-1}\left(I - W_1\left(W_1^T(X_2D_2X_2^T)^{-1}W_1\right)^{-1}W_1^T(X_2D_2X_2^T)^{-1}\right) \\ &X_2L_2X_2^TW_2 = \lambda W_2,\end{aligned} \quad (25)$$

where I is a unit matrix.(25)is equivalent to (19). $W_A, W_B, X_BL_BX_B^T, X_BD_BX_B^T$ replace $W_1, W_2, X_2L_2X_2^T, X_2D_2X_2^T$ respectively.(18)is converted into (17).

Similarly, we get W_B through (26):

$$\begin{aligned}&P_BW_B = \lambda W_B, \\ &P_B = (X_BD_BX_B^T)^{-1} \\ &*\left(I - W_A\left(W_A^T(X_BD_BX_B^T)^{-1}W_A\right)^{-1}W_A^T(X_BD_BX_B^T)^{-1}\right) \\ &*X_BL_BX_B^T,\end{aligned} \quad (26)$$

In order to eliminate correlation of discriminant transform between data set C and data set A, and between data set C and data set B, we construct the objective function and constraint:

$$\max_{W_C} J(W_C) = \frac{\left|W_C^TX_CL_CX_C^TW_C\right|}{\left|W_C^TX_CD_CX_C^TW_C\right|}, \quad (27)$$

$$s.t. \quad W_C^TW_A = 0, W_C^TW_B = 0$$

$X_C = [x_{C1}, x_{C2}, \cdots, x_{CN}]$ is training samples of dataset C, $W = [W_A, W_B], W_C, X_CL_CX_C^T, X_CD_CX_C^T$ replace $W_1, W_2, X_2L_2X_2^T, X_2D_2X_2^T$ respectively.(28)is converted into (27).

Similarly, we get W_B through (28):

$$\begin{aligned}&P_CW_C = \lambda W_C, \\ &P_C = (X_CD_CX_C^T)^{-1} \\ &*\left(I - W\left(W^T(X_CD_CX_C^T)^{-1}W\right)^{-1}W^T(X_CD_CX_C^T)^{-1}\right) \\ &*X_CL_CX_C^T,\end{aligned} \quad (28)$$

where $W = [W_A, W_B]$, I is a unit matrix. W_C is a matrix that consists of d_C eigenvectors corresponding to d_C different nonzero eigenvalues of P.

C. Algorithm description

The brief of our approach is given as follows:

Step 1: Calculate W_A, W_B, W_C by using (2),(12),(14)on training set A, B and C, respectively.

Step 2: Project all samples on $W_A^{'} = [W_A; W_B; W_C]$, then classify the test sample with the nearest neighbor classifier.

Step 3: Calculate W_B, W_C in the same way, then classify the test sample with the cosine distance classifier.

III. Experiments

Our experiments are carried out on the AR, FERET and CAS-PEAL face database.

The AR face database contains over 4000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, under different lighting conditions and with various occlusions. Most of the pictures were taken in two sessions (separated by two weeks). Each session yielded 13 color images, with 119 individuals (65men and 54 women) participating in each session. We selected images from 119 individuals for use in our experiment for a total number of 3094 (=119 × 26) samples. All color images are transformed into gray images and each image was scaled to 60×60 with 256 gray levels. Fig. 1 illustrates all of the samples of one subject.

The FERET face database contains 2200 face images belonging to 200 persons, and there are 11 images corresponding to each person. Each image is 384 ×256 with 256 gray levels. Considering many images in this database include the background and the body chest region, we automatically cropped every image sample. We scaled the intercepted images to 60 × 48. Fig. 2 shows 11 images of an individual of the FERET face database.

The CAS-PEAL face database we employed contains 1060 images of 106 individuals (10 images each person) with varying lighting. A frontal image of each subject was captured under variable illumination. In the experiment, each image was automatically cropped and scaled to 60 × 48. Fig. 3 shows 10 images of an individual of the CAS-PEAL face database.

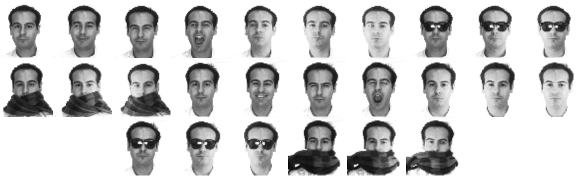

Figure 1. Demo images of one subject from the AR face database.

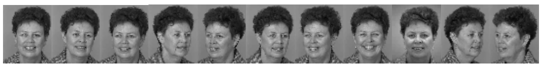

Figure 2. Demo images of one subject from the FERET face database.

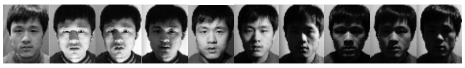

Figure 3. Demo images of one subject from the CAS-PEAL face database.

We scaled the intercepted images to 60× 48 on three face data set. Our supervised and unsupervised approach are carried out on the AR，FERET and CAS-PEAL face database and the training samples range from 2 to 6. Experimental results on these three public face databases demonstrate that the proposed approach acquires higher recognition rates when comes to small sample size.

From Table1 and Table2, we can see that the training samples range from 2 to 6, and the recognition rate of every training sample of our approach high than that of LDA on three face dataset. Our approach improves recognition rate when the number of labeled training samples is very small.

TABLE I. RECOGNITION RATES(%) OF LDA AND OUR APPROACH ON THREE DATASET

		Number of training samples				
		2	***3***	***4***	***5***	***6***
AR	***LDA***	68.00	68.32	69.52	70.43	77.69
	Our Approach	69.22	73.62	74.03	76.71	82.52
FERET	***LDA***	42.33	48.88	51.50	61.50	54.40
	Our Approach	47.83	57.19	64.00	81.75	75.60
CAS-PEAL	***LDA***	75.00	88.68	91.20	91.89	92.45
	Our Approach	75.23	89.26	92.04	92.16	92.75

TABLE II. RECOGNITION RATES(%) OF LPP AND OUR APPROACH ON THREE DATASET

		Number of training samples				
		2	***3***	***4***	***5***	***6***
AR	***LDA***	66.11	67.81	68.83	69.03	75.13
	Our Approach	69.96	71.79	73.53	74.47	79.87
FERET	***LDA***	29.44	34.16	35.93	38.67	33.80
	Our Approach	48.33	55.50	59.79	77.17	71.90
CAS-PEAL	***LDA***	61.67	59.16	63.68	75.47	74.06
	Our Approach	63.44	65.90	77.67	78.87	84.91

IV. CONCLUSIONS

The experiments demonstrate our proposed image feature extraction method based on orthogonal projection transformation of multi-task learning technology improves the recognition rate relative to LDA and LPP. When the number of labeled training samples is very small, the sample information we can use would be very little and the recognition rates of traditional image recognition methods are not satisfactory. Our approach apply the useful information contained in other databases to help extract the features more effectively and improve the recognition rates.

ACKNOWLEDGEMENTS

The work described in this paper was fully supported by the NSFC under Project No. 61073113 and Project No. 61272273, the New Century Excellent Talents of Education Ministry under Project No. NCET-09-0162, the Doctoral Foundation of Education Ministry under Project No. 20093223110001.

REFERENCES

[1] B. Bakker and T. Heskes, “Task clustering and gating for Bayesian multi-task learning,” Journal of Machine Learning Research, vol. 4, pp. 83-99, 2003.

[2] T. Heskes, “Empirical Bayes for learning to learn,” Proceeding of ICML-2000, ed. Langley, P., pp. 367-374, 2000.

[3] L.F. Chen, H.Y.M. Liao, M.T. Ko, G.J. Yu, “A New LDA-based Face Recognition System Which Can Solve the Small Sample Size Problem,” Pattern Recognition, vol. 33, no. 1, pp. 1713-1726, 2000.

[4] Q.X. Gao, L. Zhang, D. Zhang, “Face Recognition Using FLDA with Single Training Image Per-person,” Applied Mathematics and Computation, vol. 205 no.12, pp. 726-734, 2008.

[5] X Y Jing, Zhang D and Jin Z, “UODV: improved algorithm and generalized theory,” Pattern Recognition, vol. 36, pp. 2593-2602, 2003.

[6] R. Caruana, “Multitask learning,” Journal of Machine. Learning, vol. 1, pp. 41-75, 1997.

[7] R. Caruana, “Learning Many Related Tasks at the Same Time with Backpropagation,” NIPS, pp. 657-664, 1995.

[8] J. Ghosn and Y. Bengio, “Bias Learning, Knowledge Sharing,” IEEE Trans. Neural Networks, vol. 14, no. 4, pp. 748-765, 2003.

[9] D. Silver and R. Mercer, “Selective Functional Transfer: Inductive Bias From Related Tasks,” In Proc. IASTED Int. Conference on AI, Soft Comput, pp. 182-189, 2001.

[10] D. Silver and R. Mercer, “The Task Rehearsal Method of Life-long Learning: Overcoming Impoverished Data,” Canadian Conference on AI, pp. 90-101, 2002.

[11] A. Argyriou, T. Evgeniou and M. Pontil. “Multi-task Feature Learning,” NIPS, pp. 41-48, 2006.

[12] T. Jebara, “Multi-task Feature and Kernel Selection for SVMs,” International Conference on Machine Learning, July, 2004.

[13] G. Obozinski, B. Taskar and M. I. Jordan, “Multi-task Feature Selection,” Technical report, Department of Statistics, University of California, Berkeley, 2006.

[14] X Y Jing, Q Liu, C Lan, J Y Man, S Li, Zhang D., “Holistic Orthogonal Analysis of Discriminant Transforms for Color Face Recognition,” Proc. Int. Conf. Image Proc, pp. 3841-3844, 2010.