IEICE TRANS. INF. & SYST., VOL.E105–D, NO.1 JANUARY 2022

LETTER

A Novel Transferable Sparse Regression Method for Cross-Database Facial Expression Recognition*

Wenjing ZHANG[†], *Nonmember*, **Peng SONG**[†a)], *Member*, *and* **Wenming ZHENG**[††], *Nonmember*

SUMMARY In this letter, we propose a novel transferable sparse regression (TSR) method, for cross-database facial expression recognition (FER). In TSR, we firstly present a novel regression function to regress the data into a latent representation space instead of a strict binary label space. To further alleviate the influence of outliers and overfitting, we impose a row sparsity constraint on the regression term. And a pairwise relation term is introduced to guide the feature transfer learning. Secondly, we design a global graph to transfer knowledge, which can well preserve the cross-database manifold structure. Moreover, we introduce a low-rank constraint on the graph regularization term to uncover additional structural information. Finally, several experiments are conducted on three popular facial expression databases, and the results validate that the proposed TSR method is superior to other non-deep and deep transfer learning methods.

key words: *sparse regression, transfer learning, cross-database facial expression recognition*

1. Introduction

Facial expression contains a large amount of emotional information, which plays an important role in verbal and nonverbal communications. Facial expression recognition (FER) aims to classify facial expressions into the following emotional states, e.g., anger, fear, disgust, happiness, sadness, and surprise. With the rapid development of artificial intelligence, FER has attracted much attention in many application fields, e.g., human-computer interaction, educational tutoring systems, and pain detection.

Recently, regression based methods are very popular to improve the FER performance [1]–[3]. For example, in [1], Wang et al. propose an unsupervised feature selection method for FER, which is based on spectral regression and manifold learning. In [2], Yan et al. present a regression based robust locality preserving projection (RRLPP) method, which can effectively eliminate the noises and occlusion in FER. In [3], Peng et al. develop a low-rank spectral regression (LRSR) model, which decomposes the projection matrix in spectral regression by two-factor matrices to conduct subspace learning. Note that all these methods try to regress the original data into the label space, which can achieve promising performance. However, there exist two main shortcomings. First, in practice, the training and testing data may be sampled from different scenes, e.g., different lighting, equipment, and environment. Second, directly transforming the original data into a binary label space is too strict and may cause overfitting. Thus, the recognition rate may drop significantly for cross-database FER.

To solve the above-mentioned problems, some label relaxation transfer learning methods have been proposed. For example, in [4], Xu et al. propose a discriminative transfer subspace learning (DTSL) method. DTSL introduces a label relaxation matrix, and imposes low-rank and sparse constraints on the reconstruction matrix to achieve better transfer performance. In [5], Zhang et al. present a guide subspace learning (GSL) method for transfer learning, which also introduces a label relaxation matrix to improve the discriminative of subspace. In [6], Chen et al. develop a robust transferable subspace learning (RTSL) for cross-corpus FER, which jointly considers the distance divergence and label relaxation guidance strategy to transfer knowledge.

Motivated by the above discussions, in this letter, we propose a novel transferable sparse regression (TSR) method to solve the cross-database FER problem. Different from the aforementioned methods, the proposed TSR approach firstly regresses the source data into a latent representation space. Meanwhile, we constrain the regression term by using an $\ell_{2,1}$-norm, which can reduce the noise and outliers. Then, we introduce a pairwise relation constraint to further exploit the discriminative information. In addition, we develop a novel low-rank constrained global graph, which not only can preserve the geometric structure of the cross-database data, but also can uncover the structural information.

2. Proposed Method

Here we first elaborate the definitions of terminologies, which are frequently used in this letter. For the cross-database FER tasks, let $X_s \in R^{d\times n_s}$ and $X_t \in R^{d\times n_t}$ be the source and target emotional data, respectively, where d is the dimensionality of the original data, n_s and n_t indicate the numbers of source and target samples, respectively. Let

Manuscript received July 15, 2021.
Manuscript revised September 9, 2021.
Manuscript publicized October 12, 2021.

[†]The authors are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China.
[††]The author is with the Key Laboratory of Child Development and Learning Science, Ministry of Education (Southeast University), and Research Center for Learning Science, Southeast University, Nanjing 210096, China.

*This work is partially supported by the National Natural Science Foundation of China under Grant 61703360, the Fundamental Research Funds for the Central Universities under Grants 2242021k30014 and 2242021k30059, and the Graduate Innovation Foundation of Yantai University (GIFYTU).

a) E-mail: pengsong@ytu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2021EDL8062

$W \in R^{d\times c}$ be the projection matrix, where c is the dimensionality of the common subspace. Define $V \in R^{n_s\times c}$ as the latent representation matrix. $Y \in R^{n_s\times c}$ as the binary label matrix of the source database. For an arbitrary matrix $A \in \mathbb{R}^{n_s\times n_t}$, the nuclear norm of A is defined as $\|A\|_* = \sum_i \sigma_i(A)$, where $\sigma_i(A)$ is the singular values of A, and the $\ell_{2,1}$-norm of A is defined as $\|A\|_{2,1} = \sum_{i=1}^{n_s} \sqrt{\sum_{j=1}^{n_t} A_{ij}^2}$.

Note that there is rich information hidden in the source database, e.g., the label discriminative information and the inherent geometric structural information. How to effectively utilize this valuable information will contribute to learn a more robust subspace and boost the recognition performance. As is known to all, linear regression (LR) can give full consideration to the label discriminative information. However, conventional LR algorithms assume that the training samples will be transformed into a strict binary label space, or learn more transformation matrices to relax the strict label matrix. Although they can achieve satisfactory results, learning multiple transformation matrices will be very time consuming.

Instead of directly regressing the training data into the label space, in this letter, we firstly regress the training data into a latent representation space V, which is defined as

$$\min_W \|V - X_s^T W\|_F^2 \tag{1}$$

It can be noticed that Eq. (1) is constrained by a Frobenius norm, which is sensitive to noises and outliers. To address this problem, as [7], we rewrite Eq. (1) as follows:

$$\min_W \|V - X_s^T W\|_{2,1} \tag{2}$$

Then, we introduce a pairwise constraint to further exploit the label discriminative information. It assumes that two samples in the latent representation space are close, only if the samples with the same category are close to each other. Therefore, the pairwise constraint can be reformulated as minimizing the following distance-distance difference problem:

$$\min_V \|V^T V - Y^T Y\|_F^2 \tag{3}$$

Since manifold learning has recently been widely applied on transfer subspace learning [6], we further take into account the cross-database local geometric structural information. In this letter, we construct a global k-nearest neighbor graph to preserve the geometric structure information and achieve knowledge transfer, which can be formulated as

$$\min_W tr(W^T X L X^T W) \tag{4}$$

where $L = D - W$ is a Laplacian matrix, D is a diagonal matrix, and its diagonal entry $D_{ii} = \sum_{j\neq i} W_{ij}$. The weight matrix W is a binary weighting matrix that defines the similarity of each pair of samples, which is defined as

$$W_{ij} = \begin{cases} 1, & x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Because the Laplacian matrix L is a real symmetric, we use the eigen-decomposition technique on Eq. (4). Then the following equation can be obtained as

$$\begin{aligned} tr(W^T X L X^T W) &= tr(W^T X U S U^T X^T W) \\ &= \|S^{\frac{1}{2}} U^T X^T\|_F^2 = \|AW\|_F^2, \end{aligned} \tag{6}$$

where $A = S^{\frac{1}{2}} U^T X^T$.

As for the cross-database FER, the low-rank constraint [4] can not only achieve more effective representation, but also can reveal the additional subspace structure. Therefore, we impose the nuclear norm on the graph constraint, and rewrite the Eq. (6) as

$$\min_W \|AW\|_* \tag{7}$$

By integrating Eqs. (2), (3) and (7), the final objective function can be formulated as

$$\begin{aligned} &\min_{V,W} \|V - X_s^T W\|_{2,1} + \alpha\|V^T V - Y^T Y\|_F^2 + \gamma\|AW\|_* \\ &s.t. \quad V^T \mathbf{1} = \mathbf{1},\ W^T W = I. \end{aligned} \tag{8}$$

It is obvious that Eq. (8) is non-convex, which is hard to directly be solved. Therefore, in this letter, we use the alternating direction method of multipliers (ADMM) [8] to solve this problem. Firstly, we introduce two auxiliary variables G and H to make the objective function separable. And Eq. (2) can be expressed as

$$\|V - X_s^T W\|_{2,1} = 2Tr(V - X_s^T W)^T Q (V - X_s^T W) \tag{9}$$

For clarity, we set $F = V - X_s^T W$, $Q = [q_{ii}]$ is a diagonal matrix with $q_{ii} = \frac{1}{2\|F^i\|^2+\epsilon}$, ϵ is a small positive constant to avoid dividing by zero. Then we solve Eq. (8) by minimizing the following Lagrangian function:

$$\begin{aligned} \mathcal{L} = {} & Tr(V - X_s^T W)^T Q (V - X_s^T W) + \alpha\|V^T H - Y^T Y\|_F^2 + \gamma\|G\|_* \\ & + <Y_1, V - H> + <Y_2, G - AW> + tr\left(\phi(I - W^T W)\right) \\ & + \frac{\mu}{2}\left(\|V - H\|_F^2 + \|G - AW\|_F^2\right) \end{aligned} \tag{10}$$

where Y_1 and Y_2 are the Lagrange multipliers, α, γ and ϕ are the trade-off parameters, and $\mu > 0$ is a penalty parameter. The detailed procedures of solving (10) are given as follows.

1) **Update V by fixing the other variables.** Let the partial derivative of $\mathcal{L}$ with respect to V equal 0, we can obtain

$$V^* = (\mu I + \alpha H H^T + Q)^{-1}(Q X_s^T W + \mu H - Y_1 + \alpha H Y^T Y) \tag{11}$$

Since the column normalization constraint $\|V_{:,i}\|_{i=1}^{n_s} = \mathbf{1}$, the optimal V^* is

$$V^* = [V_{:,1}^*, V_{:,2}^*, \ldots, V_{:,n}^*] \tag{12}$$

2) **Update H by fixing the other variables.** Let the partial derivative of $\mathcal{L}$ with respect to H equal 0, we can obtain

$$H^* = (\alpha VV^T + \mu I)^{-1}(\alpha VY^T Y + \mu V + Y_1) \quad (13)$$

3) **Update G by fixing the other variables**, we can obtain

$$\arg\min_G \gamma\|G\|_* + <Y_2, G - AW> + \frac{\mu}{2}\|G - AW\|_F^2. \quad (14)$$

We utilize the singular value thresholding (SVT) [9] algorithm to solve the Eq. (14).

4) **Update W by fixing the other variables.** Let the partial derivative of $\mathcal{L}$ with respect to W equal 0, we can obtain

$$W^* = (X_s Q X_s^T + \phi W + A^T A)^{-1}(X_s Q V + A^T G + \frac{Y_1}{\mu} A^T) \quad (15)$$

5) **Update the multipliers Y_1, Y_2 and μ:**

$$\begin{cases} Y_1 = Y_1 + \mu(V - H) \\ Y_2 = Y_2 + \mu(G - AW) \\ \mu = \min(\rho\mu, \mu_{max}) \end{cases} \quad (16)$$

where $\rho > 1$ and μ_{max} are the constants.

3. Experiments

To evaluate the performance of the proposed TSR approach, we conduct extensive cross-database and non cross-database FER experiments on three commonly used emotional databases, including JAFFE[†], CK+[††], and KDEF[†††]. And we compare the proposed TSR approach with the following transfer learning methods, including principal component analysis (PCA) [10], transfer component analysis (TCA) [11], transfer joint matching (TJM) [12], discriminative transfer subspace learning (DTSL) [4], joint geometrical and statistical alignment (JGSA) [13], domain invariant and class discriminative (DICD) [14], and guide subspace learning (GSL) [5]. In addition, we also compare our method with several deep transfer learning methods, i.e., deep domain confusion (DDC) [15], and deep adaptation networks (DAN) [16].

In our experiments, for a fair comparison, we firstly resize all facial images into 60×60 pixels and transform them into grayscale. Then we use the local binary patterns (LBP) feature for feature extraction. In this way, each emotional image is divided into nine regions, and we can obtain a 2304 dimensional feature. In addition, we also use a fine-tuned ResNet-50 model [17] to further extract the deep features of facial images, and the dimensionality of each image is set to 2048. We choose six common emotion categories for evaluation, including anger, disgust, fear, happiness, sadness, and surprise. And we conduct six types of cross-database FER experiments (source → target), including J → C, J → K, C → J, C → K, K → C, K → J, where J, C and K are abbreviations for JAFFE, CK+ and KDEF, respectively. As for the non cross-database FER experiments, we divide each database into six parts by categories, then randomly select 7/10 for training, and the rest are used for testing. The experiments are repeated five times and the average recognition results are given.

For fairness, we strictly follow the same experimental settings, and empirically search the optimal parameter values for evaluation. Then, we give the best experimental results of each method. Specifically, the subspace dimension of PCA, TCA, TJM, JGSA and DICD is set to 100, and the dimensionality of DTSL, GSL and the proposed TSR method is set to six. The main three trade-off parameters, i.e., α, γ and ϕ, are involved in our method, which are tuned from $[0.001, 0.01, 0.1, 1, 10, 100]$. Also, we set $\mu = 0.001$, $\eta = 0.001$, the number of nearest neighbors $k = 30$, and the maximum iteration number $T = 10$. To evaluate the recognition performance, a linear support vector machine (SVM) is used as the baseline classifier for all the compared methods.

The recognition accuracy of different methods using LBP feature and deep features are reported in Tables 1 and 2, respectively. From the tables, we can obtain the following observations.

From Table 1, we can observe that, the proposed approach can obtain better recognition results compared with other classical transfer learning methods. Under the six experimental settings, the average recognition accuracy of the proposed TSR is 55.82%, which obtains a 2.61% improvement compared with the best baseline method GSL. We can find that the proposed method can achieve the highest performance in all cross-database FER tasks. Also, we can obtain that, the recognition rate about considering label regression methods, such as DTSL, GSL, and our proposed TSR, is higher than other transfer learning methods. These results all demonstrate that, by considering the sparse regression and low-rank graph constraint, the proposed method can obtain a more robust subspace for cross-database FER.

We further give the results of different methods using deep features in Table 2. From the table, we can notice that, by using the deep features, the performance of the proposed TSR method also outperforms that of all the non-deep transfer learning methods. When comparing the performance between deep features and LBP feature, it can be found that, the accuracies of deep features are much better than those of LBP feature. The reason may be that the deep features can obtain deeper and better structural information, which is helpful for facial expression classification. Furthermore, by comparing TSR with deep transfer learning algorithms, i.e., DDC and DAN, the proposed TSR method also achieves better recognition performance, with about 5.54% improvement in average recognition accuracy.

In order to provide fairer evaluation of the proposed method against the other methods, we also give the results of non cross-database FER in Table 3. From the table, we can notice that, the performance of the proposed TSR method is better than all the compared methods on all databases.

[†]http://www.kasrl.org/jaffe.html
[††]http://www.pitt.edu/emotion/ck-spread.htm
[†††]http://www.emotionlab.se/kdef/

Table 1 Recognition performance (%) of different methods using the LBP feature under different settings.

Settings	Compared methods							
	PCA	TCA	TJM	DTSL	JGSA	DICD	GSL	TSR
J→C	41.43	37.62	51.90	49.52	47.14	49.52	54.29	**56.19**
J→K	47.14	40.95	51.67	52.86	51.43	50.00	50.95	**55.24**
C→J	37.70	36.61	36.61	44.81	37.70	42.08	43.17	**45.90**
C→K	55.24	51.90	55.71	55.71	56.19	54.76	57.62	**59.05**
K→C	58.10	60.48	64.29	63.81	65.62	63.81	65.71	**67.14**
K→J	46.45	42.08	46.99	45.36	49.73	43.72	50.27	**51.37**
Average	47.95	44.94	51.20	52.01	49.61	50.65	53.21	**55.82**

Table 2 Recognition performance (%) of different methods using the deep features under different settings.

Settings	Compared methods									
	PCA	TCA	TJM	DTSL	JGSA	DICD	GSL	DDC*	DAN*	TSR
J→C	43.33	45.71	51.90	58.57	51.90	52.38	55.24	58.57	65.24	**66.19**
J→K	56.67	57.62	63.33	66.19	60.95	69.52	65.24	56.19	67.14	**72.38**
C→J	46.45	43.17	51.91	55.74	53.01	55.19	51.91	53.55	50.82	**56.28**
C→K	62.38	61.9	66.67	70.48	66.19	68.57	70.74	69.05	67.62	**74.76**
K→C	61.43	62.86	67.14	69.52	64.76	68.10	66.19	71.43	63.81	**73.33**
K→J	55.74	54.64	59.02	62.30	59.02	**63.93**	62.84	49.73	58.47	63.39
Average	54.33	54.32	60.00	63.80	59.31	62.95	62.03	59.75	62.18	**67.72**

Table 3 Recognition performance (%) using LBP and deep features under different databases.

Features	Databases	Compared methods							
		PCA	TCA	TJM	DTSL	JGSA	DICD	GSL	TSR
LBP	JAFFE	60.00	65.97	68.42	69.82	73.68	75.09	76.14	**77.54**
	CK+	64.26	68.37	70.82	75.08	78.69	79.35	81.97	**83.28**
	KDEF	63.67	66.67	65.52	68.50	75.00	75.04	74.66	**79.33**
Deep features	JAFFE	85.96	89.82	90.53	88.42	90.17	92.28	92.63	**95.08**
	CK+	86.89	87.87	90.49	90.82	91.47	89.51	91.15	**94.10**
	KDEF	88.00	89.40	91.00	90.34	90.67	90.33	91.34	**92.67**

This demonstrates that the proposed method can not only effectively improve the recognition performance of cross-database FER, but also can be applied to non cross-database FER tasks.

4. Conclusion

In this letter, we have presented a novel transfer learning method, called transferable sparse regression (TSR), for cross-database FER. Specifically, we develop a sparse regression function to regress the data into a latent subspace. Meanwhile, we utilize a pairwise relation term to guide feature transfer learning. In addition, we introduce a low-rank graph constraint to uncover the local geometric structure of cross-database data. Extensive experiments on three public facial expression databases demonstrate the superiority of the proposed method.

References

[1] L. Wang, K. Wang, and R. Li, "Unsupervised feature selection based on spectral regression from manifold learning for facial expression recognition," IET Computer Vision, vol.9, no.5, pp.655–662, 2015.
[2] J. Yan, B. Yan, R. Liang, G. Lu, H. Li, and S. Xie, "Facial expression recognition via regression-based robust locality preserving projections," IEICE Trans. Inf. & Syst., vol.E101-D, no.2, pp.564–567, 2018.
[3] Y. Peng, L. Zhang, W. Kong, F. Qin, and J. Zhang, "Low rank spectral regression via matrix factorization for efficient subspace learning," Journal of Intelligent & Fuzzy Systems, vol.39, no.3, pp.3401–3412, 2020.
[4] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation," IEEE Trans. Image Process., vol.25, no.2, pp.850–863, 2016.
[5] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, and C.P. Chen,

"Guide subspace learning for unsupervised domain adaptation," IEEE Trans. Neural Netw. Learn. Syst., vol.31, no.9, pp.3374–3388, 2019.

[6] D. Chen, P. Song, W. Zhang, W. Zhang, B. Xu, and X. Zhou, "Robust transferable subspace learning for cross-corpus facial expression recognition," IEICE Trans. Inf. & Syst., vol.E103-D, no.10, pp.2241–2245, 2020.

[7] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," Advances in Neural Information Processing Systems, vol.23, 2010.

[8] S. Boyd, N. Parikh, and E. Chu, Distributed optimization and statistical learning via the alternating direction method of multipliers, Now Publishers, 2011.

[9] J.-F. Cai, E.J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM Journal on optimization, vol.20, no.4, pp.1956–1982, 2010.

[10] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol.2, no.1-3, pp.37–52, 1987.

[11] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," IEEE Trans. Neural Netw., vol.22, no.2, pp.199–210, 2011.

[12] M. Long, J. Wang, G. Ding, J. Sun, and P.S. Yu, "Transfer joint matching for unsupervised domain adaptation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1410–1417, 2014.

[13] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1859–1867, 2017.

[14] S. Li, S. Song, G. Huang, Z. Ding, and C. Wu, "Domain invariant and class discriminative feature learning for visual domain adaptation," IEEE Trans. Image Process., vol.27, no.9, pp.4260–4273, 2018.

[15] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," arXiv preprint arXiv:1412.3474, 2014.

[16] M. Long, Y. Cao, Z. Cao, J. Wang, and M.I. Jordan, "Transferable representation learning with deep adaptation networks," IEEE Trans. Pattern Anal. Mach. Intell., vol.41, no.12, pp.3071–3085, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.770–778, 2016.