Interspeech 2022
18-22 September 2022, Incheon, Korea

Coupled Discriminant Subspace Alignment for Cross-database Speech Emotion Recognition

Shaokai Li[1], Peng Song[1], Keke Zhao[1], Wenjing Zhang[1], Wenming Zheng[2]*

[1]School of Computer and Control Engineering, Yantai University, Yantai, China
[2]Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
pengsong@ytu.edu.cn

Abstract

Speech emotion recognition (SER) is a long-standing important research problem in speech signal processing. In practice, the training and test data are often collected in different scenarios, e.g., different languages, different collecting devices, which would severely degrade the recognition performance. To tackle this problem, in this paper, we propose a novel transfer learning algorithm, named coupled discriminant subspace alignment (CDSA), for cross-database SER. In CDSA, we first conduct linear discriminant analysis (LDA) in source and target databases, respectively. Meanwhile, we learn a latent common subspace, where the target samples are represented by the combination of source samples. Furthermore, we align the projection subspace of source and target databases to make the model more robust. Extensive experiments are carried out on four benchmark databases, and the results demonstrate the effectiveness of the proposed method.

Index Terms: linear discriminant analysis, transfer learning, coupled projection, speech emotion recognition

1. Introduction

The goal of speech emotion recognition (SER) is to identify the corresponding emotion categories from speech signals, e.g., happiness, anger, sadness, fear, disgust, and surprise [1]. In recent years, it has been shown impressive performance in various applications [2], e.g., safety driving assist system, automatic translation, and assistant diagnostic tool in medical treatment.

In SER tasks, feature extraction and feature classification are two important parts. Among which the feature extraction refers to extracting the emotional features from speech signals, while feature classification refers to training a classification model using the extracted features [3]. Over the past decades, many classification algorithms have been employed for SER [4, 5, 6, 7]. These algorithms are carried out on the assumption that the training and test data are from the same database, and follow similar distributions, which cannot be satisfied in practical scenarios due to the difference in gender, age and recording scenes.

Recently, transfer learning has shown appealing performance in handling the above-mentioned mismatch problems [8, 9]. The mechanism of transfer learning is to transfer the knowledge gained from training data to test data by utilizing the distance metric strategies. As an important distance metric in transfer learning, maximum mean discrepancy (MMD) [10], has been widely used [11]. For example, in [12], Long et al. adopt MMD to minimize both the marginal and conditional probability distribution to obtain a common feature representation. Recently, Zhang et al. combine MMD with linear discriminant analysis (LDA), and develop a coupled projection to align the subspace while retaining the discriminant information of the source database [13]. Zong et al. propose a database regeneration label space (DRLS) method [14] for the cross-database micro-expression problem. In [15], Li et al. integrate MMD and manifold learning to deal with the domain adaptation problem.

Over the past decade, many scholars have tried to develop transfer learning algorithms for cross-database SER. For example, in [16], Hassan et al. have introduced three transfer learning algorithms for cross-database SER. In [17], Deng et al. develop a database-adaptive auto-encoder approach to learn the database- invariant features. In [18, 19], Zong et al. present a regression-based algorithm for cross-database SER. More recently, in [20], Song et al. develop a transfer linear subspace learning framework for cross-database SER. In [21], Zhang et al. present a joint transfer subspace learning and regression (JTSLR) method for SER.

The above-mentioned algorithms can alleviate the "database bias" problem to some extent. However, they do not consider the specific property of each database, which is important for knowledge transfer [9]. Therefore, in this paper, we propose a novel transfer subspace learning algorithm, named coupled discriminant subspace alignment (CDSA), for cross-database SER. The basic idea of CDSA is to retain the shared discriminant information of source and target databases in the process of knowledge transferring. We simultaneously exploit the specific and common subspace by learning the coupled discriminant subspace and the sparse reconstruction term. We further reduce the coupled projection subspace discrepancy to improve the transfer performance. The flowchart of CDSA is shown in Fig. 1.

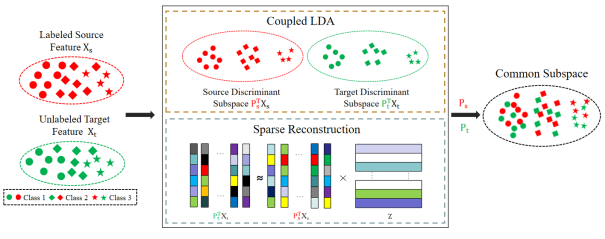


Figure 1: *The framework of CDSA. The red color represents the source database and the green color represents the target database, and different shapes represent different categories.*

*Corresponding author.

10.21437/Interspeech.2022-40

2. The proposed method

We begin with an introduction of the main notations used in this work. Denote that $X_s \in \mathbf{R}^{d\times n_s}$ is the labeled source feature matrix and $X_t \in \mathbf{R}^{d\times n_t}$ is the unlabeled target feature matrix, and n_s and n_t are the corresponding numbers of samples and d represents the dimension of features. $L_s \in \mathbf{R}^{d\times d}$ and $L_t \in \mathbf{R}^{d\times d}$ are the scatter matrices of source and target databases, respectively, $P_s \in \mathbf{R}^{d\times d}$ and $P_t \in \mathbf{R}^{d\times d}$ are the source and target projection matrices, respectively, and $Z \in \mathbf{R}^{n_s\times n_t}$ is the sparse reconstruction matrix.

2.1. The objective function

We first learn coupled projection matrices to obtain the discriminant information of the source and target databases, respectively. Note that LDA requires the guidance of real labels, and the target database is unlabeled. Thus, here we utilize the virtual labels which are learned by a linear support vector machine (SVM). As discussed in [22], conventional LDA cannot well deal with the small sampled size problem. To tackle this problem, Li et al. present a maximum margin criterion approach [23], in which the margin between different classes is maximized. Thus, the objective function of the proposed coupled LDA is written as follows:

$$\begin{aligned}&\min_{P_s,P_t} \mathrm{Tr}(P_s^T L_s P_s) + \mathrm{Tr}(P_t^T L_t P_t)\\ &\text{s.t. } P_s^T P_s = I,\ P_t^T P_t = I\end{aligned} \tag{1}$$

where $\mathrm{Tr}(\cdot)$ means the trace of a matrix, $L_s = S_w^s - \mu S_b^s$, $L_t = S_w^t - \mu S_b^t$, in which $S_w^s \in \mathbf{R}^{d\times d}$ and $S_b^s \in \mathbf{R}^{d\times d}$ are the within-class and between-class scatter matrices of the source database, respectively, and $S_w^t \in \mathbf{R}^{d\times d}$ and $S_b^t \in \mathbf{R}^{d\times d}$ are the within-class and between-class scatter matrices of the target database, respectively, and μ is a constant with small value. The constraints $P_s^T P_s = I$ and $P_t^T P_t = I$ are used to avoid calculating the inverse of scatter matrices.

Then, we adopt a linear reconstruction strategy in the learned common subspace, in which all target samples are represented by the combination of source samples. Hence the divergence between these two databases can be reduced. In addition, we impose an $\ell_{2,1}$-norm constraint on the sparse reconstruction matrix. The objective function is written as follows:

$$\min_{P_s,P_t,Z} \|P_s^T X_s Z - P_t^T X_t\|_F^2 + \gamma\|Z\|_{2,1} \tag{2}$$

where $||\cdot||_F^2$ and $||\cdot||_{2,1}$ are the Frobenius norm and $\ell_{2,1}$-norm, respectively, Z is a sparse reconstruction matrix, and γ is a regularization parameter.

To further reduce the divergence between two databases, as [24], we utilize a simple but efficient algorithm to align the dual projection subspace, and the objective function is written as

$$\min_{P_s,P_t} \|P_s - P_t\|_F^2 \tag{3}$$

Combining Eqs. (1) (2) and (3), the common and specific information between two databases can be well exploited, and we can get the final objective function as follows:

$$\begin{aligned}\min_{P_s,P_t,Z}\ &\mathrm{Tr}(P_s^T L_s P_s) + \mathrm{Tr}(P_t^T L_t P_t) + \beta\|P_s - P_t\|_F^2\\ &+ \alpha\|P_s^T X_s Z - P_t^T X_t\|_F^2 + \gamma\|Z\|_{2,1}\\ \text{s.t. } &P_s^T P_s = I,\ P_t^T P_t = I\end{aligned} \tag{4}$$

where α and β are positive trade-off parameters.

Algorithm 1 The CDSA algorithm

Input: The labeled source feature matrix X_s and unlabeled target feature matrix X_t; the label matrix Y_s of source database; the regularization parameters α, β, γ; and a small threshold value ε.
Output: P_s, P_t.
Initialize: Initialize P_t via PCA; Initialize L_s, L_t and set $t = 0$.
repeat
1. Fix other variables and update P_s by using Eq. (7);
2. Fix other variables and update P_t by using Eq. (9);
3. Fix other variables and update Z by using Eq. (12);
4. Update the target pseudo labels via SVM;
5. Update the scatter matrix L_t;
6. $t = t + 1$;
7. Check the convergence conditions: $\Delta T = T^{(t)} - T^{(t-1)} < \varepsilon$, where $T^{(i)}$ represents the objective value in the i-th iteration.

until Convergence
return P_s, P_t

2.2. Optimization

In this subsection, we put forward an iterative algorithm to solve the objective function in Eq. (4). Eq. (4) can be rewritten as the following Lagrange form:

$$\begin{aligned}\mathcal{L} = &\mathrm{Tr}(P_s^T L_s P_s) + \mathrm{Tr}(P_t^T L_t P_t)\\ &+ \alpha\mathrm{Tr}\big((P_s^T X_s Z - P_t^T X_t)^T (P_s^T X_s Z - P_t^T X_t)\big)\\ &+ \beta\mathrm{Tr}\big((P_s - P_t)^T (P_s - P_t)\big) + \gamma\|Z\|_{2,1}\end{aligned} \tag{5}$$

The detail optimization procedures are listed as follows:

1) Update P_s: Update P_s by fixing the other variables, we take the derivative of $\mathcal{L}$ w.r.t. P_s, and set it to zero, we can get the closed-form solution as

$$\begin{aligned}\frac{\partial\mathcal{L}(P_s)}{\partial P_s} = &L_s P_s + \alpha X_s Z Z^T X_s^T P_s - \alpha X_s Z X_t^T P_t\\ &+ \beta P_t - \beta P_s = 0\end{aligned} \tag{6}$$

$$P_s = (L_s + \alpha X_s Z Z^T X_s^T - \beta I)^{-1}(\alpha X_s Z X_t^T P_t - \beta P_t) \tag{7}$$

2) Update P_t: Update P_t by fixing the other variables, we take the derivative of $\mathcal{L}$ w.r.t. P_s, and set it to zero, we can obtain its closed-form solution as

$$\begin{aligned}\frac{\partial\mathcal{L}(P_t)}{\partial P_t} = &L_t P_t + \alpha X_t X_t^T P_t - \alpha X_t Z^T X_s^T P_s\\ &+ \beta P_t - \beta P_s = 0\end{aligned} \tag{8}$$

$$P_t = (L_t + \alpha X_t X_t^T + \beta I)^{-1}(\alpha X_t Z^T X_s^T P_s + \beta P_s) \tag{9}$$

3) Update Z: Due to $\|Z\|_{2,1}$ is not smooth, we first calculate its sub-gradient matrix $Q \in \mathbf{R}^{n_s\times n_s}$ [25]:

$$Q_{ii} = \begin{cases} 0, & \text{if } z^i = 0\\ \frac{1}{2\|z^i\|}, & \text{otherwise}\end{cases} \tag{10}$$

where z^i is the $i-$th row of the matrix Z. By fixing the matrix Q, we take the derivative of $\mathcal{L}$ w.r.t. Z, which is written as

$$\frac{\partial\mathcal{L}(Z)}{\partial Z} = \alpha X_s^T P_s P_s^T X_s Z + \gamma Q Z - \alpha X_s^T P_s P_t^T X_t \tag{11}$$

Table 1: *Recognition accuracy* (%) *on different tasks. The best performance is shown in boldface.*

Tasks	Traditional methods		Transfer learning methods							Ours
	LDA	SDA	DRLS	TCA	JDA	DaLSR	JTSLR	JGSA	LPJT	
E→e	41.26	32.54	40.08	43.25	**42.46**	42.06	42.06	36.50	40.47	**42.46**
E→R	25.00	24.78	32.78	32.77	33.88	31.11	33.89	38.33	42.77	**43.88**
E→B	40.38	25.00	44.25	33.58	34.90	34.62	46.15	36.53	**50.00**	**50.00**
e→E	32.74	39.82	41.59	38.93	39.82	49.56	43.36	41.59	**52.21**	46.90
e→R	25.55	42.78	28.33	45.55	41.66	42.22	45.00	34.44	39.44	**46.11**
e→B	32.69	34.62	42.31	38.43	38.57	42.31	42.23	44.23	44.23	**48.07**
R→E	24.77	35.00	36.28	41.59	45.13	36.28	47.79	51.09	49.55	**51.32**
R→e	24.20	33.73	35.71	**39.68**	36.50	30.76	31.75	31.76	37.69	38.49
R→B	27.69	28.85	27.31	30.76	32.69	32.69	32.69	36.53	34.61	**40.38**
B→E	32.74	28.32	49.56	44.11	45.58	51.33	48.04	**61.94**	53.09	50.44
B→e	28.26	31.52	34.29	31.02	31.90	34.24	**36.61**	34.78	29.89	35.05
B→R	25.55	29.44	28.33	38.88	40.55	40.56	36.11	39.44	33.33	**44.44**
Average	30.06	32.19	36.73	38.21	38.63	38.97	40.47	40.59	42.27	**44.79**

Setting the above equation to zero, we can obtain the solution for Z as

$$Z = (\alpha X_s^T P_s P_s^T X_s + \gamma Q)^{-1}(\alpha X_s^T P_s P_t^T X_t) \qquad (12)$$

The procedures of CDSA are summarized in Algorithm 1.

2.3. Complexity

In this subsection, we give the complexity analysis of the proposed CDSA. For computing P_s, according to Eq. (7), the complexity is $O(d^3 + d^2 n_s^2 n_t + d^3 n_s n_t)$. For computing P_t, according to Eq. (9), the computational complexity is $O(d^3 + d^2 n_t + d^3 n_s n_t)$. For computing Z, according to Eq. (12), the complexity is $O(d^3 n_s^2 + n_s^3 + d^3 n_s n_t)$. To sum up, the total computational complexity is about $O\big(T(d^3 + d^2 n_s^2 n_t + d^3 n_s n_t + d^2 n_t + d^2 n_s^2 + n_s^3)\big)$, where T is the number of iterations.

3. Experiment

3.1. Experimental setup

In this subsection, we evaluate the performance of the proposed algorithm for cross-database SER. Four public benchmark databases, including EmoDB (E) [26], eNTERFACE'05 (e) [27], BAUM-1a (B) [28], and RML (R) [29], are employed in our experiments. Two of the above databases are randomly selected as the source and target databases, respectively, and 12 groups of cross-database SER tasks (source database→target database: E→e, E→R, E→B, e→E, e→R, e→B, R→E, R→e, R→B, B→E, B→e, B→R), are conducted. We select five common emotional categories, i.e., anger (AN), sadness (SA), disgust (DI), happiness (HA), and fear (FE), in our experiments. All the source database is selected, and the target database is randomly divided into 10 parts, among which 7/10 samples are used for training and the remainder are used for testing. To ensure the reliability of the experimental results, the experiments are repeated 10 times and the average results are reported.

The proposed CDSA is compared with several state-of-the-art subspace learning and transfer subspace learning methods, including LDA, semi-supervised discriminant analysis (SDA) [30], transfer component analysis (TCA) [31], joint distribution adaptation (JDA) [12], domain-adaptive least squares regression (DaLSR) [18], joint geometrical and statistical alignment (JGSA) [13], domain regeneration in the label space (DRLS) [14], locality preserving joint transfer (LPJT) [15], and joint transfer subspace learning and regression (JTSLR) [21]. We choose the linear SVM as the baseline classifier, and use the recognition accuracy of the test database for evaluation, which is written as

$$\text{accuracy} = \frac{|x : x \in D_{test} \wedge \widehat{y}(x) = y(x)|}{|x : x \in D_{test}|} \qquad (13)$$

where D_{test} is the test database, $y(x)$ is the true label of x, and $\widehat{y}(x)$ is the predicted label.

In the experiments, we use the openSMILE toolkit [32] to extract the 1582-dimensional low-level features using the standard feature set used in INTERSPEECH 2010 paralinguistic challenge [33]. For the settings of hyper-parameters, since the training and the testing data follow different probability distributions, we cannot directly adopt the cross-validation to determine the values of parameters [34]. Thus, we search the optimal values in the parameter space $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$.

3.2. Results and Analysis

The recognition results are shown in Table 1. From the table, we have the following findings.

Firstly, it can be found that the proposed CDSA method significantly achieves better recognition accuracy than the nine compared methods. The average recognition accuracy of the proposed method is higher than the best baseline LPJT with 2.52% improvement. These results show that the proposed method can learn more robust representation for cross-database SER tasks.

Secondly, it can be observed that the transfer learning methods significantly outperform the traditional subspace learning methods, i.e., LDA and SDA. The limitation of these two approaches lies in that they do not consider the divergence across

the source and target databases. On the contrary, the transfer learning methods can effectively address this shortcoming and obtain better recognition performance.

Thirdly, CDSA significantly outperforms LDA, SDA and JTSLR, which are the discriminate subspace learning approaches. The reason can be attributed to that, only a single subspace projection is not enough for cross-database tasks when the divergence across databases is very large. CDSA can well address this problem by developing a novel coupled subspace projection.

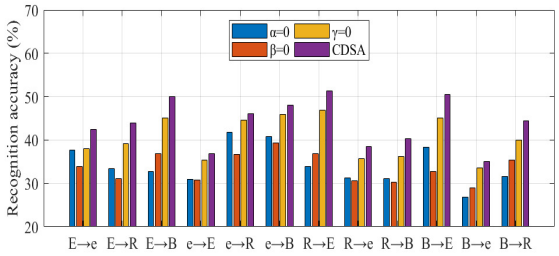
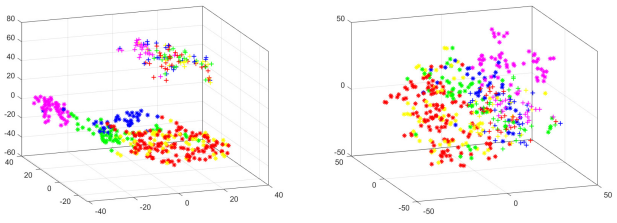


Figure 2: *Ablation study of CDSA under 12 tasks.*



(a) Original data (b) CDSA

Figure 3: *t-SNE visualization of data representation on the E→B task. The asterisk points and the cross points represent the source and target samples, respectively, and different colors mean different emotion categories.*

3.3. Ablation study and visualization of data

In this subsection, we give the ablation study of CDSA. We analyze the effectiveness by considering the following aspects, i.e. linear reconstruction, coupled projection subspace alignment, and sparse reconstruction matrix. The results are given in Fig. 2. From the figure, we have the following observations.

- Firstly, when $\alpha = 0$ in Eq. (4), the linear reconstruction term is ignored, resulting in a significant decline in the recognition accuracy. This result proves that the linear reconstruction term plays a positive role in our method.
- Secondly, when $\beta = 0$ in Eq. (4), the coupled projection alignment term is ignored, the recognition accuracy on all cases drops significantly. This proves that this term also plays a positive role in our method.
- Thirdly, when $\gamma = 0$ in Eq. (4), the recognition accuracy also decreases, but the impact on the recognition accuracy is not significant.

Also, we give the t-SNE [35] visualization results of the proposed method on the E→B task. Fig. 3 (a) shows that there exists a gap between the original feature distribution of source and target databases. Fig. 3 (b) shows the projected data by the proposed method. From the figure, we can find that, after projection, both the source and target databases follow similar feature distribution, and the samples from the same category are close to each other.

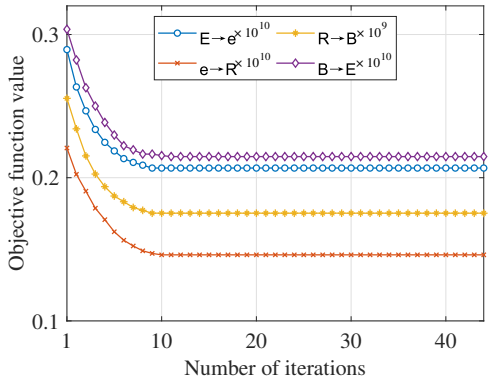


Figure 4: *Convergence analysis of CDSA.*

3.4. Convergence analysis

Since the objective function is solved by an iterative algorithm, the speed of convergence is important for the efficiency of the proposed method. In this subsection, we empirically check the convergence of the proposed CDSA. Here, we choose four representative groups of tasks, i.e., E→e, e→R, R→B, and B→E, covering the four databases used in the experiments. Fig. 4 plots the convergence curves of CDSA on these four tasks. As can be seen from the figure, the proposed method on all settings converges quickly, and usually converges within about 10 iterations, which validates the effectiveness of CDSA.

4. Conclusions

In this paper, we present a novel coupled discriminant subspace alignment (CDSA) approach for cross-database SER. This method extends traditional LDA to a transferable manner, so that the divergence across different databases can be reduced significantly. It first performs discriminate subspace learning in each database separately. Then, a linear reconstruction regularization in the learned subspace is developed to reduce the divergence across databases. Furthermore, the coupled projection subspace is aligned to make the model more robust. Extensive experiments are carried out on four benchmark databases, and the results show that the proposed CDSA achieves superior performance than state-of-the-art compared algorithms. In the future, we will investigate to extract effective deep features, and integrate the proposed method into the deep transfer learning framework to obtain better recognition results.

5. Acknowledgements

This work was supported in part 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).

6. References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[3] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. IEEE, 2003, pp. II–1.

[5] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[6] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.

[7] X. Wu, Y. Cao, H. Lu, S. Liu, D. Wang, Z. Wu, X. Liu, and H. M. Meng, "Speech emotion recognition using sequential capsule networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[8] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[9] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.

[10] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[11] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[12] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2200–2207.

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

[14] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2484–2498, 2018.

[15] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6103–6115, 2019.

[16] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[17] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[18] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, 2016.

[19] J. Zhang, L. Jiang, Y. Zong, W. Zheng, and L. Zhao, "Cross-corpus speech emotion recognition using joint distribution adaptive regression," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3790–3794.

[20] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition." *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, 2019.

[21] W. Zhang, P. Song, D. Chen, C. Sheng, and W. Zhang, "Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[22] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of LDA," in *Object Recognition Supported by User Interaction for Service Robots*, vol. 3. IEEE, 2002, pp. 29–32.

[23] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, 2006.

[24] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2960–2967.

[25] J. Li, J. Zhao, and K. Lu, "Joint feature selection and structure preservation for domain adaptation." in *IjCAI*, 2016, pp. 1697–1703.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[27] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[28] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[29] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2016.

[30] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proceeding of the 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–7.

[31] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[33] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

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

[35] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2559–2566.