CCBR 2023 第十七届中国生物特征识别大会
The 17th Chinese Conference on Biometric Recognition

Dynamic Graph-Guided Transferable Regression for Cross-Domain Speech Emotion Recognition

Shenjie Jiang[1], Peng Song[1*], Run Wang[1], Shaokai Li[1], and Wenming Zheng[2]

[1]Yantai University

[2]Southeast University

CONTENTS:

01 Background

01 Background

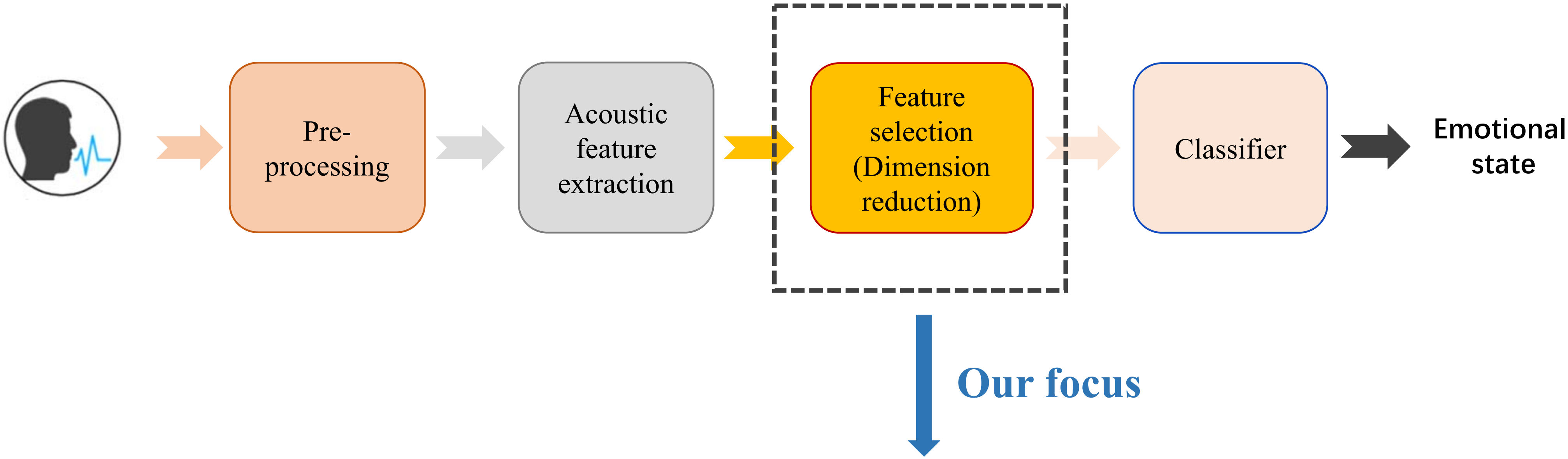
The main purpose of **Speech Emotion Recognition (SER)** is to classify speech signals according to different emotions, such as anger, disgust, fear, happiness, and sadness. It is widely used in various popular fields such as affective computing, pattern recognition, signal processing and human-computer interaction.

Driving assist system

Automatic translation

Robot interaction

01 The process of SER



Learning a transfer subspace, which can obtain a common subspace to reduce the discrepancy between databases.

01 Traditional SER method

Many classification algorithms have been employed for SER，including：

- Hidden Markov model (HMM)
- Gaussian mixture model (GMM)
- Support vector machine (SVM)
- Neural network (NN)
- Deep neural network (DNN)
- Sparse representation
- Regression algorithms

02 The challenging problem

02 The challenging problem

- **Data distribution mismatch problem:** in practical application scenarios, the speaker's gender, language, age and so on are different.
- **Underutilize Label Information:** The label information in the source domain has not been fully utilized to guide the transfer.

02 Transfer learning

Transfer learning: The idea of transfer learning is to transfer the knowledge gained from one domain (source domain) to learn the knowledge of related but different domain (target domain).

We take the labeled database as the source domain and the unlabeled database as the target domain. The transfer learning can be used to solve the cross-domain SER problem.

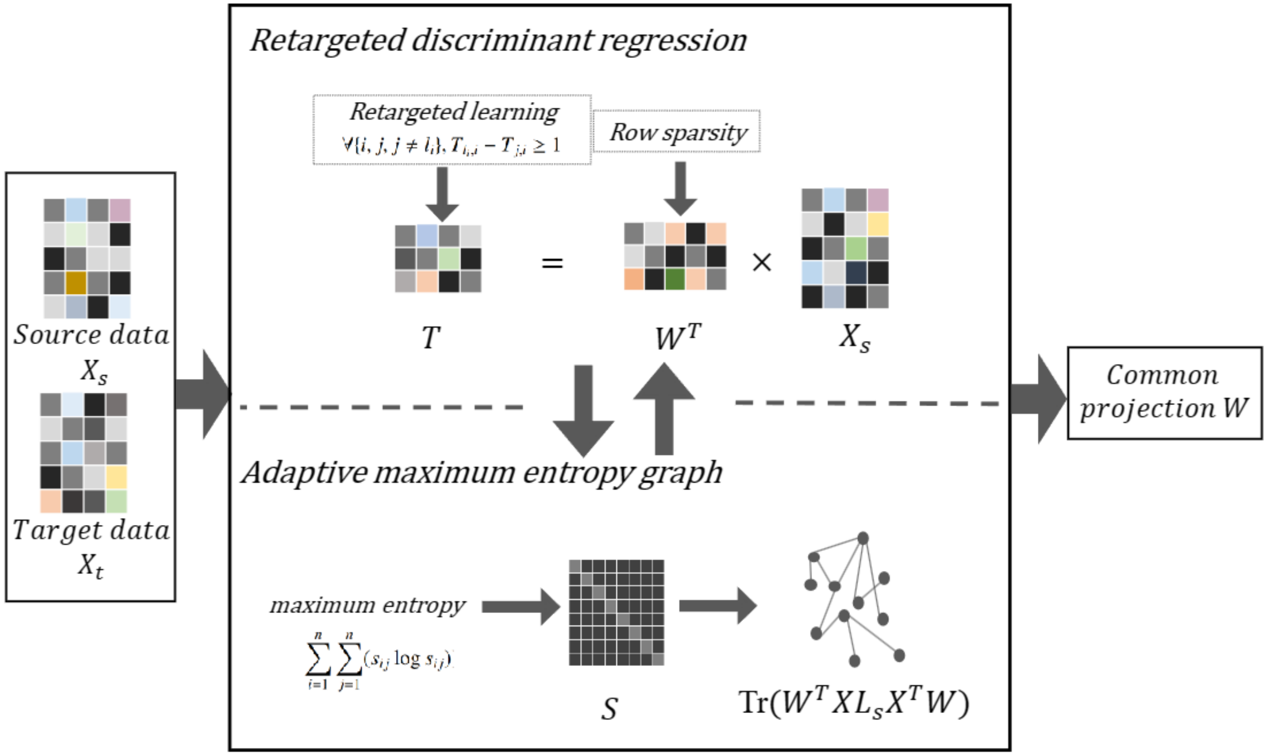
02 The related works

Transfer learning for cross-domain SER:

- transfer component analysis (TCA) 2010
- joint distribution adaptation (JDA) 2013
- transfer joint matching (TJM) 2014
- balanced distribution adaptation (BDA) 2017
- transfer linear discriminant analysis (TLDA) 2018
- robust discriminative sparse regression (RDSR) 2020
- joint transfer subspace learning and regression (JTSLR) 2021
- transferable discriminant linear regression (TDLR) 2022
- unsupervised transfer components learning (UTCL) 2023

03 The proposed method

03 Our method framework



03 The discriminant regression

Discriminative regression is a classic approach commonly used in classification tasks. To address the inherent trade-off between model flexibility and overfitting, we rewrite traditional discriminative regression to efficiently utilize label information of the source domain.

$$\min_{W,T} \|T - W^T X_s\|_F^2 + \gamma \|W\|_{2,1}$$

$$s.t. \forall \{i, j, j \neq l_i\}, T_{l_i,i} - T_{j.i} \geq 1$$

03 The dynamic graph regularization

During the projection from high-dimensional space to low-dimensional subspace, the inherent local geometric structure of data may be destroyed. To address this issue, we introduce the graph Laplacian. And we introduce an adaptive learning strategy into the process of transfer learning, which can learn an adaptive manifold structure by adaptively updating the similarity matrix. With the following formula, the distribution gap between the two domains can be effectively minimized.

$$\min_{W,S} Tr(W^T X L_S X^T W) + \sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij} log s_{ij})$$

$$s.t. W^T W = I, \sum_{j=1}^{n} s_{ij} = 1, s_{ij} > 0$$

03 Our method DGTR

Combining the above two equations, the objective function of DGTR is formulated as follows:

discriminant regression

$$\min_{W,T,S} \|T - W^T X_s\|_F^2 + 2\alpha\left(Tr(W^T X L_S X^T W) + \beta \sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij} log s_{ij})\right) + \gamma \|W\|_{2,1}$$

$$s.t. \forall\{i,j,j \neq l_i\}, T_{l_i,i} - T_{j.i} \geq 1, W^T W = I, \sum_{j=1}^{n} s_{ij} = 1, s_{ij} > 0$$

sparse constraint

dynamic graph regularization

04 Experiments

04 Experimental setup

Databases: Berlin (B), IEMOCAP (I), CVE (C), and TESS (T).

Emotional categories: anger (AN), neutral (NE), happiness (HA), and sadness (SA)

Feature Extraction: For low-level features, we use the 1582-dimensional standard feature set in INTERSPEECH 2010 Paralinguistic challenge and use the linear support vector machine (**SVM**) as the baseline classifier.

For deep features, we use ResNet-50 model on Mel spectrograms to obtain 2048-dimensional deep features.

Tasks: 12 cross-domain SER tasks, i.e., C→B, I→B, T→B, B→C, I→C, T→C, B→I, C→I, T→I, B→T, C→T, and I→T.

04 Experimental results

Table 1. Recognition accuracy (%) of different algorithms using low-level features.

Settings	Compared methods								DGTR
	LDA	TCA	JDA	TJM	BDA	TLDA	JTSLR	TDLR	
C→B	60.22	65.98	60.82	67.01	57.27	59.79	66.72	69.83	**71.13**
I→B	50.14	50.52	53.61	53.61	**59.21**	56.41	52.73	52.58	56.70
T→B	54.67	55.43	51.97	57.66	56.41	54.21	55.77	**58.59**	56.89
B→C	58.62	53.21	48.51	48.08	50.41	55.56	50.76	57.69	**67.87**
I→C	40.79	40.38	51.28	41.03	49.32	**54.49**	46.17	49.63	48.08
T→C	52.65	54.37	55.46	52.45	53.21	56.72	58.13	56.46	**62.18**
B→I	42.28	43.73	37.42	43.21	48.52	32.44	44.21	41.45	**50.85**
C→I	40.71	46.77	46.77	47.29	44.10	50.19	48.13	50.59	**50.97**
T→I	38.66	44.23	40.92	46.89	47.53	44.23	47.55	42.67	**49.56**
B→T	52.85	55.52	56.33	53.59	**63.78**	54.87	55.78	53.22	56.50
C→T	55.41	54.56	55.95	56.66	55.61	53.21	**58.73**	57.88	58.58
I→T	50.11	55.33	50.40	50.16	51.87	52.54	54.73	51.48	**59.38**
Average	49.76	51.67	50.79	51.47	53.10	52.06	53.28	53.51	**57.39**

04 Experimental results

Table 2. Recognition accuracy (%) of different algorithms using deep features.

Settings	Compared methods							DGTR
	JDA	BDA	JTSLR	DAAN*	MRAN*	DSAN*	BNM*	
C→B	58.76	56.71	53.61	67.90	**70.68**	65.43	42.90	70.10
I→B	57.73	56.83	63.92	65.12	66.05	**67.59**	60.19	61.86
T→B	63.92	53.83	58.76	**66.05**	55.25	61.73	43.83	55.67
B→C	56.41	56.71	47.44	38.58	45.30	47.03	60.27	**71.79**
I→C	51.17	46.85	49.36	40.31	45.49	48.75	45.30	**51.28**
T→C	57.69	51.78	49.36	43.57	44.34	59.88	41.65	**62.26**
B→I	47.36	32.58	48.85	48.29	48.24	**48.98**	45.55	46.40
C→I	**43.88**	40.44	43.65	38.78	40.62	40.20	30.00	43.21
T→I	45.81	**45.98**	41.79	41.07	39.55	38.91	36.26	45.88
B→T	51.67	46.83	69.71	57.16	53.35	**70.11**	46.65	54.38
C→T	51.25	57.10	**62.71**	52.10	58.47	57.79	42.40	57.08
I→T	**56.25**	47.81	53.75	43.90	45.78	55.10	47.28	51.88
Average	53.49	49.45	53.58	50.23	51.09	55.12	45.19	**55.82**

04 Experimental results

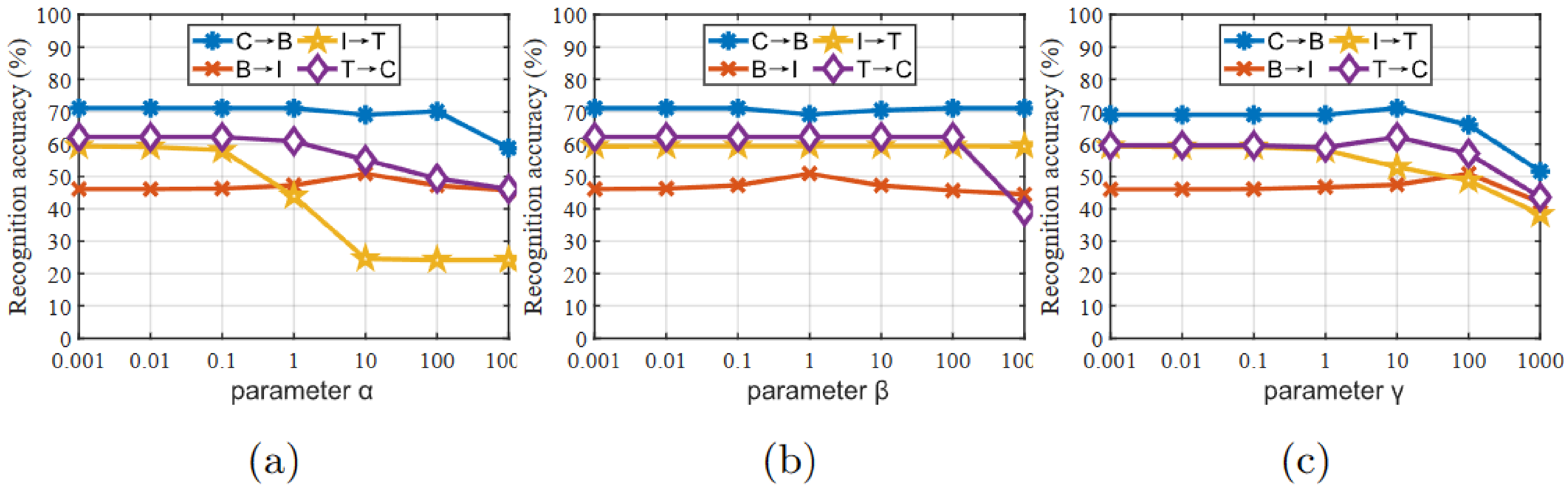
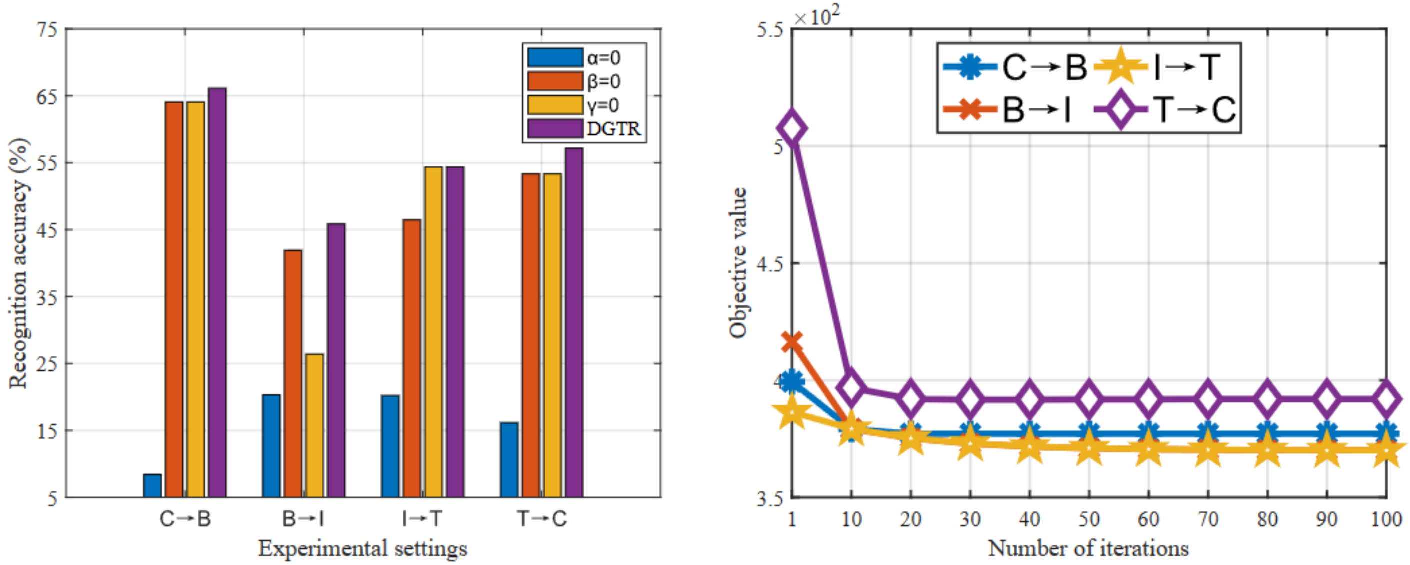


Fig. 1. Parameter sensitivity of the proposed DGTR w.r.t. (a) α, (b) β, and (c) γ.

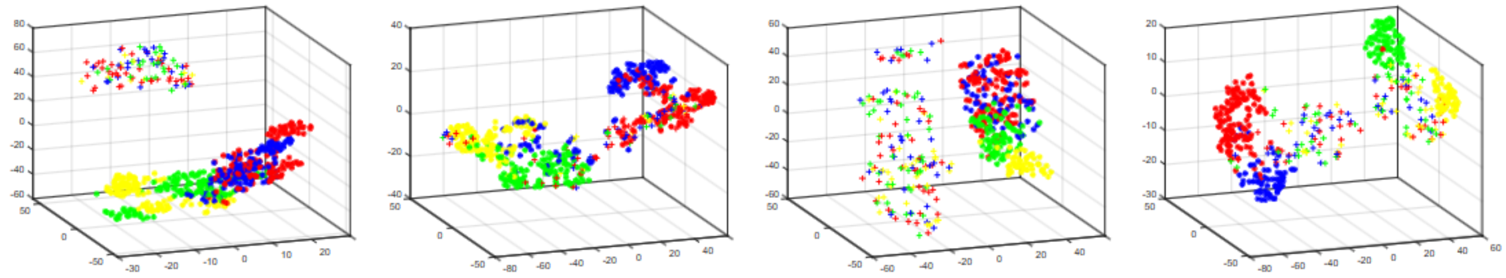
04 Experimental results



(a) Ablation analysis (b) Convergence curves

Fig. 2. Ablation analysis and convergence curves of DGTR.

04 Experimental results



(a) Original data (b) Our method (c) Original data (d) Our method

Fig. 3. t-SNE visualization on the tasks C→B (the first two figures) and B→C (the last two figures). The “∗” and “+” indicate the source and target data, respectively.

Conclusion:

We propose a novel cross-domain SER method, named dynamic graph-guided transferable regression (DGTR). It utilizes the source labels to guide the procedures of transfer, and designs a dynamic graph to effectively minimize the distribution gap across two domains. We also impose an $\ell 2,1$-norm on the projection matrix to make the model robust. Experimental results show the superiority of DGTR over some state-of-the art methods.

In the future, we will integrate the proposed method into the deep transfer learning framework to obtain better recognition results.

Thank You!