SUPER WIDE REGRESSION NETWORK FOR UNSUPERVISED CROSS-DATABASE FACIAL EXPRESSION RECOGNITION

Na Liu[1,4,2], *Baofeng Zhang*[2,1], *Yuan Zong*[3], *Li Liu*[4],*Jie Chen*[4], *Guoying Zhao*[4], *Junchao Zhu*[2*]

[1]School of Computer Science and Engineering, Tianjin University of Technology, China
[2]School of Electrical and Electronic Engineering, Tianjin University of Technology, China
[3]Research Center for Learning Science, Southeast University, China
[4]Center for Machine Vision and Signal Analysis, University of Oulu, Finland

ABSTRACT

Unsupervised cross-database facial expression recognition (FER) is a challenging problem, in which the training and testing samples belong to different facial expression databases. For this reason, the training (source) and testing (target) facial expression samples would have different feature distributions and hence the performance of lots of existing FER methods may decrease. To solve this problem, in this paper we propose a novel super wide regression network (SWiRN) model, which serves as the regression parameter to bridge the original feature space and the label space and herein in each layer the maximum mean discrepancy (MMD) criterion is used to enforce the source and target facial expression samples to share the same or similar feature distributions. Consequently, the learned SWiRN is able to predict the expression categories of the target samples although we have no access to any label information of target samples. We conduct extensive cross-database FER experiments on CK+, eNTERFACE, and Oulu-CASIA VIS facial expression databases to evaluate the proposed SWiRN. Experimental results show that our SWiRN model achieves more promising performance than recent proposed cross-database emotion recognition methods.

Index Terms— Cross-database facial expression recognition, transfer learning, domain adaptation, super wide network

1. INTRODUCTION

In the past few decades, facial expression recognition (FER) has become a very active research topic among affective computing, pattern recognition, and human-computer interaction (HCI) [1]. Generally speaking, the aim of FER is to learn a classifier based on a set of labeled training facial expression samples such that the learnt classifier can accurately predict the expression categories, e.g., happiness, fear, anger, sadness, surprise and disgust, of the unlabeled testing samples [2]. Currently, many researchers have been devoted to investigating FER [3, 4] and proposed various methods. It is notable that among the evaluation of the existing FER methods, the training and testing samples usually belong to the same facial expression database and hence the training and testing samples can be thought to share the same or similar feature distribution. However, in many practical scenarios, the training (source) and testing (target) facial expression samples may come from two different databases. In this case, the original similar feature distribution between the training and testing facial expression samples would not be satisfied. Consequently, the performance of the above mentioned well-performed FER methods may decrease significantly. This thus brings us a challenging but interesting topic, i.e., cross-database FER.

As a typical domain adaptation (DA) [5] problem, cross-database FER can follow the categorization of DA. According to the provided label information of the target (testing) facial expression samples, cross-database can be roughly divided to two cases including semi-supervised case and unsupervised case [6, 7]. In semi-supervised case, the label information of target facial expression samples is available, while in unsupervised case, all the target facial expression samples are completely unknown. In this paper, we will focus on unsupervised case in cross-database facial expression recognition problem, which is more difficult and has drawn lots of attention of researchers in recent years. To solve this challenging problem, Sangineto et al. [8] proposed a transductive parameter transfer method to obtain satisfactory parameters of target-specific classifiers, which is transferred from the parameter information provided by the source-specific classifiers. In the work of [9], Zheng et al. proposed a novel transductive transfer subspace learning (TTSL) framework to deal with unsupervised cross-database FER problem. Their TTSL framework aims at learning a discriminative subspace to predict the expression categories of the unlabelled target samples by combining labelled source samples and an auxiliary set that is selected from target facial expression samples. Recently, a domain

*Corresponding author

adaptive dictionary learning (DADL) method is proposed for unsupervised cross-database FER problem, aiming to learn a common dictionary [6]. Besides the above methods, it is also worth mentioning the work of selective transfer machine (STM) [10, 11], which is proposed for personalized facial action unit detection. STM inherits the ability of kernel mean matching (KMM) [12] to eliminate the feature distribution difference between source and target samples and also have the discriminative ability of support vector machine (SVM).

In this paper, we propose a simple but effective deep learning model called super wide regression network (SWiRN) to handle the unsupervised cross-database FER problem. We illustrate the basic idea of SWiRN and its structure in Fig. 1. As Fig. 1 shows, we use a super wide network (two fully connected layers in this paper) to serve as the regression parameter instead of using projection matrix in subspace learning to build the relationship between the facial expression features and labels. Meanwhile, by using MMD [13] criterion as regularization, we enforce the output of SWiRN with source and target samples as input, respectively, to have the same or similar feature distribution.

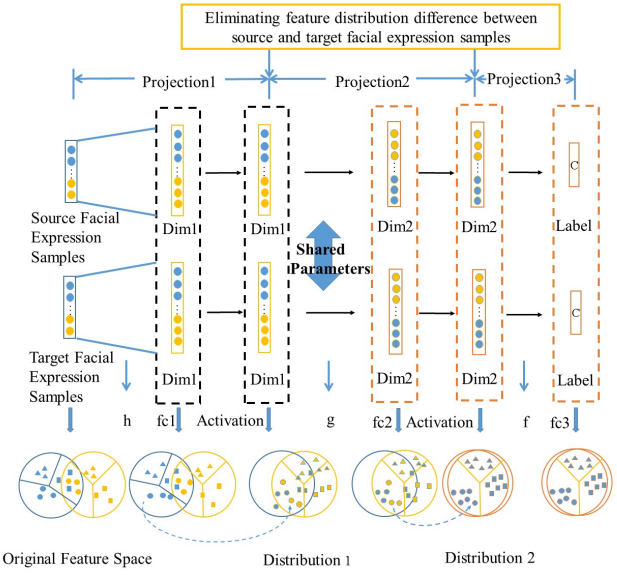

Fig. 1. Structure of the Super Wide Regression Network for Unsupervised Cross-Database Facial Expression Recognition

2. CROSS-DATABASE FER USING SUPER WIDE REGRESSION NETWORK

2.1. Notations

In this section, we first introduce the notations which will be used in this paper. Suppose that $\mathbf{X}_s \in \mathbb{R}^{d\times n_s}$ and $X_t \in \mathbb{R}^{d\times n_t}$ are the feature matrices extracted from the faical expression samples belonging to the source and target databases, respectively, where d is the dimension of the expression feature and n_s and n_t denote the numbers of source and target facial expression samples. According to the problem setting of unsupervised cross-database facial expression recognition, only the label information of source database is known. Therefore, let $\mathbf{L}_s \in \mathbb{R}^{c\times n_s}$ be the label matrix of source facial expression samples, where c is the number of facial expressions. Each column $\mathbf{l}_i^s$ $(i = 1, \cdots, n_s)$ of $\mathbf{L}_s$ is called the class vector whose elements are binary, i.e., 0 or 1. Only the j^{th} element of $\mathbf{l}_i^s$ is 1 and others are all 0 if its corresponding facial sample belongs to the j^{th} facial expression.

2.2. Linear Regression Model

Regression methods [14, 1, 15] has been successfully applied to FER and other emotion recognition problems. The basic idea of regression method for FER is straightforward, i.e., learning a regression parameter to build the relationship between expression features and labels, which can be formulated as follows:

$$\min_{\mathbf{U}} \|\mathbf{L}_s - \mathbf{U}^T\mathbf{X}_s\|_F^2, \tag{1}$$

where $\mathbf{U}$ is the regression parameter. In fact, from the view of subspace learning, Eq.(1) can be also interpreted as learning a projection matrix to bridge the original expression feature space and the expression label space. Hence, we can extend this simple yet effective regression method to a domain adaptive version such that the learned projection matrix can transform the target samples from feature space to label space as well. More importantly, it is hoped in the label space, the transformed source and target samples would have the same or similar feature distributions. Thus, the regression method can be applicable to cross-database FER problem. To this end, we first design a simple criterion to measure the distribution difference in a subspace, which is derived from MMD [13]. It is known that according to the definition of MMD, the distribution distance can be viewed as the mean difference in the Hilbert space. In our case, we simply use the feature mean vector to measure the distribution distance between the source and target feature sets. More specifically, the optimal learned regression parameter $\mathbf{U}$ should also be the one, which can minimize the following optimization problem:

$$\min_{\mathbf{U}} \|\frac{1}{n_s}\mathbf{U}^T\mathbf{X}_s\mathbf{1}_s - \frac{1}{n_t}\mathbf{U}^T\mathbf{X}_t\mathbf{1}_t\|^2, \tag{2}$$

where $\mathbf{1}_s$ and $\mathbf{1}_t$ are the vectors whose elements are all ones and their dimensions are n_s and n_t, respectively. By using Eq.(2) as the regularization term for Eq.(1), we will arrive at the optimization problem as follows:

$$\min_{\mathbf{U}} \|\mathbf{L}_s - \mathbf{U}^T\mathbf{X}_s\|_F^2 + \lambda\|\frac{1}{n_s}\mathbf{U}^T\mathbf{X}_s\mathbf{1}_s - \frac{1}{n_t}\mathbf{U}^T\mathbf{X}_t\mathbf{1}_t\|^2, \tag{3}$$

2.3. Super Wide Regression Network

Recently, deep learning (DL) network such as autoencoder (AE), convolutional neural network (CNN), and recurrent neural network (RNN) have shown its effective applications in many vision tasks [16]. One major reason is that DL networks have powerful nonlinear representation abilities. For example, CNN can learn an excellent representation of images and has been widely used in various image based research fields. We will resort to the nonlinearity of DL network and leverage it to refine the regression model shown in Eq.(3). To differentiate with linear regression model in Eq.(3), we call this new regression model super wide regression network (SWiRN). Simply speaking, we use the network to serve as the regression parameter instead of the original $\mathbf{U}$ in Eq.(3). In this paper, a simple full connection network with two hidden layers are employed, whose detailed structure is shown in Fig. 1. It should be pointed out that different from subspace learning version Eq.(3), the feature mean vector based regularization like Eq.(2) is simultaneously applied on two hidden layers and hence the final optimization problem of SWiRN becomes as follows:

$$\min_{f,g,h} \mathcal{L}(\mathbf{L}_s, f(g(h(\mathbf{X}_s)))) + \lambda(\|\frac{1}{n_s}h(\mathbf{X}_s)\mathbf{1}_s - \frac{1}{n_t}h(\mathbf{X}_t)\mathbf{1}_t\|^2 + \|\frac{1}{n_s}g(h(\mathbf{X}_s))\mathbf{1}_s - \frac{1}{n_t}g(h(\mathbf{X}_t))\mathbf{1}_t\|^2),$$

where f, g, and h denote the network parameters.

Backpropagation (BP) algorithm can be used to optimize the parameters of the proposed SWiRN. Once the optimal network parameters of SWiRN are learned, we can use this SWiRN to obtain the predicted label vector with the target facial expression sample as input and further infer its corresponding facial expression category.

3. EXPERIMENTS

3.1. Databases and Experimental Protocol

In this section, we evaluate the performance of our proposed SWiRN method by conducting extensive cross-database FER experiments on three dynamic facial expression databases including CK+ [17], eNTERFACE [18], and Oulu-CASIA VIS [19]. CK+ database consists of 593 facial videos from 123 subjects, in which 327 samples are assigned one of eight basic facial expressions (Neutral, Anger, Contempt, Disgust, Fear, Happy, Sadness, and Surprise). The eNTERFACE database is composed of 1287 facial videos from 43 subjects and they are categorized into six basic expressions including Anger, Disgust, Fear, Happy, Sadness, and Surprise. Oulu-CASIA VIS database is captured in three different illumination conditions, i.e., normal, weak, and dark, by a normal visual (VIS) camera. It contains 1440 samples from 80 subjects, which are divided into six expressions. In the experiments, we choose either two of the above three databases, e.g., CK+ and eNTERFAC, to serve as source and target database, respectively. Therefore, we have totally six groups of experiments. For convenience, we denote these six experiments by Exp.1, Exp.2, $\cdots$, Exp.6, respectively, whose detailed information of source and target databases can be found from Tables 1, 2, and 3.

Local binary pattern from three orthogonal planes (LBP-TOP) [20] is employed as the expression features. According to the suggestion of the work [20], spatial division scheme is also adopted to divide the facial expression samples into a few facial blocks in the LBP-TOP extraction. In our experiments, the spatial division grid is set as 4×4. Regarding the parameters of LBP-TOP, we set the radius R as 1 and the number of neighbors P as 8, respectively. In the end, for a facial expression video sample, we use a 2832-dimensional LBP-TOP feature vector to describe it. Furthermore, we employ the weighed average recall (WAR) and the unweighted average recall (UAR) [21] as the evaluation metrics, where WAR is the normal accuracy while UAR means the accuracy per class divided by the number of classes without considerations of instances per class. Besides, for comparison purpose, we choose three typical domain adaptation methods that are successfully applied in dealing with cross-database emotion recognition, i.e., KMM+SVM [22], KLIEP+SVM [22], and STM [10, 11] to conduct the designed experiments as well. Besides, SVM without any domain adaptation is also included in the comparison. In addition, it is noted that we choose linear kernel for SVM through all the experiments and Gaussian kernel for KMM and KLIEP. To show their best performance, we use the grid search strategy for select the optimal kernel parameter for KMM and KLIEP which corresponds to the best results. For STM, there is a trade-off parameter λ. Grid search strategy is also adopted to determine the optimal λ for STM.

Finally, as to the setup of the proposed SWiRN in the experiments, we set the dimensions of two full connection layers to 10000 and 20000. Backpropagation (BP) algorithm is used to optimize the parameters of the SWiRN. The learning rate for training our network is set to $1e^{-3}$ and the parameters of SWiRN are initialized as all zeros. We record the result of each cross-database FER experiment when the optimization reaches totally 50,000 iterations and the final reported result of each experiment is the average after the experiment is ran for ten times.

3.2. Experimental Results and Discussion

The results of all the methods for different group of experiments are given in Tables 1, 2, and 3, respectively, where the best UAR and WAR in each experiment are highlighted with **bold** typeface. From the results, it is clear to see that in all the experiments, the results of the proposed SWiRN method achieves very promising results in terms of both UAR and WAR over the baseline method (SVM without any domain

Table 1. Results in terms of UAR and WAR of Exp.1 and Exp.2 cross-database FER experiments, where the common expressions (6 classes) are Angry, Disgust, Fear, Happy, Sadness and Surprise.

#	Source Database	Target Database	SVM		KMM+SVM		KLIEP+SVM		STM+SVM		SWiRN	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
1	CK+	Oulu-CASIA VIS	24.03	24.03	25.00	25.00	26.74	26.74	25.49	25.49	**28.33**	**28.33**
2	Oulu-CASIA VIS	CK+	29.52	47.57	32.68	47.57	58.67	62.78	43.07	39.48	**60.94**	**64.08**

Table 2. Results in terms of UAR and WAR of Exp.3 and Exp.4 cross-database FER experiments, where the common emotion states (6 classes) are Angry, Disgust, Fear, Happy, Sadness and Surprise.

#	Source Database	Target Database	SVM		KMM+SVM		KLIEP+SVM		STM+SVM		SWiRN	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
3	CK+	eNTERFACE	16.90	16.94	21.26	21.29	**27.06**	**23.95**	18.76	18.80	20.35	20.33
4	eNTERFACE	CK+	11.45	18.54	23.39	28.48	27.60	23.95	26.11	22.65	**30.68**	**32.20**

Table 3. Results in terms of UAR and WAR of Exp.5 and Exp.6 cross-database FER experiments, where the common emotion states (6 classes) are Angry, Disgust, Fear, Happiness/Happy, Sadness and Surprise.

#	Source Database	Target Database	SVM		KMM+SVM		KLIEP+SVM		STM+SVM		SWiRN	
			UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
5	eNTERFACE	Oulu-CASIA VIS	17.99	17.99	18.40	18.40	20.35	20.35	17.22	17.22	**26.95**	**26.95**
6	Oulu-CASIA VIS	eNTERFACE	17.67	17.72	18.99	19.04	18.37	18.41	18.06	18.10	**20.86**	**20.84**

adaptation). More importantly, we can find that in most cases including Exp.1, Exp.2, Exp.4, Exp.5, and Exp.6, the proposed SWiRN method achieves the highest UAR and WAR among all the methods. Especially in Exp.2 and Exp.4, our method get higher results (UAR 31.42%, 19.23% and WAR 16.51%, 13.66%) than the baseline (SVM). However, we notice that in Exp.1 and Exp.3, both UAR and WAR of the proposed SWiRN method achieves very small promotion (Exp.1 with UAR 4.3% WAR 4.30%, Exp.3 with UAR 3.45% WAR 3.39%), even in Exp.3 it is slightly lower than that of KMM + SVM and KLIEP + SVM. We think this is most likely due to the limited label information provided by a small number of samples in source database, which results insufficient training of our model, such as Exp.1 and Exp.3. In the experiments, the sample number of CK+ is only 309, which is much smaller than that of eNTERFACE (1287) and Oulu-CASIA VIS (1440). More unlabeled target samples may affect the discriminant ability of SWiRN since the label information given by source samples is so limited compared with a large number of unlabeled target samples.

4. CONCLUSIONS

In this paper, we have proposed a super wide regression network (SWiRN) to deal with the unsupervised cross-database FER problem. By using the super wide network to serve as the regression parameter, SWiRN method establish a relationship between the expression features and labels which is implemented in each layer of SWiRN. MMD criterion is applied to eliminate the feature distribution difference between the source and target facial expression samples. Consequently, the trained classifier based on the source samples can accurately predict the facial expression categories of the testing samples. To evaluate the proposed SWiRN based unsupervised cross-database FER method, extensive experiments are conducted on CK+, eNTERFACE and OULU-CASIA VIS databases. The experimental results demonstrate the effectiveness of the proposed SWiRN method. Since the addition of neural networks is desirable in cross-data facial expression recognition tasks, we will introduce the convolution neutral network into our SWiRN method to improve the recognition accuracy in the future.

5. ACKNOWLEDGEMENTS

This research was supported by the Natural Science Foundation of China under Grants 61172185 and 61602345, the Application Foundation and Advanced Technology Research Project of Tianjin, the Academy of Finland, Tekes Fidipro program and Infotech Oulu.

6. REFERENCES

[1] Wenming Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 71–85, 2014.

[2] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[3] Caifeng Shan, Shaogang Gong, and Peter W McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[4] Caifeng Shan, Shaogang Gong, and Peter W McOwan, "Robust facial expression recognition using local binary patterns," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*. IEEE, 2005, vol. 2, pp. II–370.

[5] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[6] Keyu Yan, Wenming Zheng, Zhen Cui, and Yuan Zong, "Cross-database facial expression recognition via unsupervised domain adaptive dictionary learning," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 427–434.

[7] Xu Zhang, Felix Xinnan Yu, Shih-Fu Chang, and Shengjin Wang, "Deep transfer network: Unsupervised domain adaptation," *arXiv preprint arXiv:1503.00591*, 2015.

[8] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 357–366.

[9] Wenming Zheng, Yuan Zong, Xiaoyan Zhou, and Minghai Xin, "Cross-domain color facial expression recognition using transductive transfer subspace learning," *IEEE Transactions on Affective Computing*, 2016.

[10] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3515–3522.

[11] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 3, pp. 529–545, 2017.

[12] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, 2008, pp. 1433–1440.

[13] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[14] Wenming Zheng and Xiaoyan Zhou, "Speech emotion recognition based on kernel reduced-rank regression," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1972–1976.

[15] Wenming Zheng, Minghai Xin, Xiaolan Wang, and Bei 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.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[18] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.

[19] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[20] Guoying Zhao and Matti Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[21] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[22] Ali Hassan, Robert Damper, and Mahesan 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.