CROSS-CULTURE MULTIMODAL EMOTION RECOGNITION WITH ADVERSARIAL LEARNING

Jingjun Liang[1], *Shizhe Chen*[1], *Jinming Zhao*[1], *Qin Jin*[1]*, *Haibo Liu*[2], *Li Lu*[2]

[1] School of Information, Renmin University of China
[2] Tencent Inc., Beijing, P.R.China

ABSTRACT

With the development of globalization, automatic emotion recognition has faced a new challenge in the multi-culture scenario - to generalize across different cultures. Previous works mainly rely on multi-cultural datasets to address the cross-culture discrepancy, which are expensive to collect. In this paper, we propose an adversarial learning framework to alleviate the culture influence on multimodal emotion recognition. We treat the emotion recognition and culture recognition as two adversarial tasks. The emotion feature embedding is trained to improve the emotion recognition but to confuse the culture recognition, so that it is more emotion-salient and culture-invariant for cross-culture emotion recognition. Our approach is applicable to both mono-culture and multi-culture emotion datasets. Extensive experiments demonstrate that the proposed method significantly outperforms previous baselines in both cross-culture and multi-culture evaluations.

Index Terms— Multimodal Emotion Recognition, Cross-culture, Adversarial Learning

1 Introduction

Automatic emotion recognition has played an important role in natural human-machine interactions, which can benefit a wide range of applications such as service industries etc [1]. With the development of globalization, world-wide users from different cultures have emerged, which has put forward a new demand for the emotion recognition system – to generalize in the multi-cultural scenario.

However, the culture factor has a non-negligible effect on the emotion recognition performance. People in different cultures tend to express emotions in different ways [2]. As shown in previous works [3, 4], in the cross-culture setting where training and testing are from different cultures, the emotion recognition performance is significantly worse than that in the mono-cultural setting where training and testing are from the same culture.

There have been a few endeavors to address the cross-culture emotion recognition problem. Chiou et al. [5] combine emotion datasets from different cultures to improve the cross-culture performance. Neumann et al. [4] utilize a small amount of target culture corpus to finetune the emotion model trained in the source culture corpus. Chen et al. [6] propose an auto-encoder to purify the emotion representation from the culture influence in a multi-task framework. However, all the above works require multi-cultural emotion datasets, which are expensive to collect especially for resource-scare cultures.

In this paper, we aim to learn an emotion-salient embedding to improve the generalization of emotion recognition in the multi-culture scenario. Basically, the emotion classification and culture classification can be considered as two conflicting objectives for learning the emotion-salient feature. The culture recognition desires to keep the culture-related information, while the emotion recognition tempts to capture the emotion-salient information that is insensitive to culture variation. Therefore, we propose a novel adversarial learning framework to take advantage of the competition between emotion-salient feature generation and culture classification, which is beneficial to alleviate the culture influence and increase the emotion salience simultaneously. Since the culture recognition is self-supervised, our approach is flexible to utilize either mono-culture or multi-culture emotion datasets for emotion recognition to achieve cross-cultural generalization.

We conduct extensive emotion recognition experiments in the Chinese and English cultures under the cross-culture and multi-culture evaluation settings. The experimental results show that the proposed adversarial learning method significantly improves the emotion recognition performance in both the cross-culture and multi-culture settings.

2 Related Works

Previous works have explored a variety of multimodal features for emotion recognition tasks. Brady et al. [7] derive high-level acoustic, visual and physiological features from the low-level descriptors using sparse coding and deep learning. Viktor et al. [8] use early fusion to concatenate multimodal features as the input for the prediction models and improves performance successfully. Chen et al.[9] propose to dynamically fuse multiple modalities at each time step. Similar to previous works, we explore emotion features from multiple modalities including the acoustic modality and facial modality in this work. In order to address the culture discrepancy,

*Corresponding author.

Hesam et al. [10] first identify the language and then select the language-specific emotion classification model.

Generative Adversarial Networks [11] was proposed using adversarial loss for generative tasks. In order to train the generator, GANs build up a discriminator to play an adversarial game with the generator, rather than setting a normal loss function for the generator. The discriminator is designed to distinguish between samples from the generator and samples from the real data while the generator learns to output samples that can fool the discriminator. Tzeng et al. [12] observes that generative modeling of input data distributions is not necessary, as the ultimate task is to learn a discriminative representation among cross domain classification task. In this case, we find that such adversarial model is able to alleviate the redundant information in classification. Rozantsev et al. [13] show that partially shared weights embedding between different domains can lead to effective adaptation in cross domain scenario. Inspired by their work, we design ours weight sharing adversarial networks based on discriminative model in this work.

3 The Proposed Approach

In this section, we firstly describe the multimodal features for emotion recognition. Then we introduce the proposed adversarial framework to learn emotion-salient and culture-invariant feature representations.

3.1 Multimodal Features

We extract two types of features from the acoustic modality and facial modality respectively.

Acoustic Modality: We utilize the open-source toolkit OpenSMILE [14] to extract the acoustic features with the configuration in INTERSPEECH 2010 [15]. The extracted feature is denoted as **IS10**, which consists of 1,582 dimensions. We apply culture-specific z-normalization on each dimension of the feature to reduce culture discrepancy in data pre-processing.

Facial Modality: We first detect faces in the video with the open-source toolkit Dlib [16]. Each face image is transformed into the gray scale with height and width of 224 pixels. Then we utilize the state-of-the-art Dense Convolutional Neural Networks [17] (DenseNet) to extract facial features. The DenseNet is pretrained on the FERPlus [18] dataset for facial expression recognition following the setup in [19]. We extract the activation from the last pooling layer of DenseNet for each face image, and apply average pooling over all faces in the video to generate the video-level facial feature, which is referred as the **Dface** feature.

3.2 Adversarial Emotion and Culture Classification

Assume $D_0 = \{(x_0^i, y_0^i)\}_{i=1}^{N_0}$ and $D_1 = \{(x_1^i, y_1^i)\}_{i=1}^{N_1}$ are two video emotion datasets in two different cultures, where x denotes the multimodal video-level feature, y denotes the emotion label which is one of the K emotional classes, and N is the size of the dataset. We omit the subscript i for notation simplicity in the following sections.

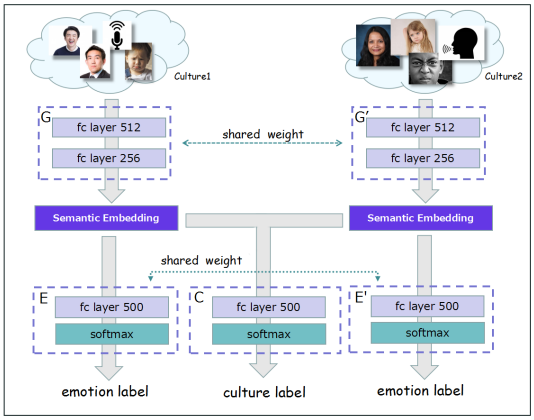


Fig. 1. Adversarial Network Structure

Our goal is to learn an emotion-salient representation z from x to generalize across cultures for emotion recognition. The proposed model is illustrated in Figure 1, which consists of three modules: feature generator (G) to transform x to z, emotion classifier (E) using z for emotion classification and culture classifier (C) using z for culture classification. We treat E and C as two adversarial targets for G, that is, G is trained to improve the classification performance in E and confuse the classification in C. Therefore, the feature z learned from G contains more emotional information rather than cultural specific information.

Specifically, we set G as a fully connected network so that $z = G(x; \theta_G)$, where θ_G is the parameter in the network. E is a neural network based classifier and the predicted emotion probability of class k is as follows:

$$\hat{y_k} = \text{softmax}(\text{E}(z; \theta_E)) \tag{1}$$

where θ_E is the parameter in E. To train the emotion classifier E, we optimize the cross-entropy loss function with the annotated emotion label y, which is:

$$L_{emo} = -\mathbb{E}_{(x_i, y_i)} \sum_{k=1}^{K} y_k log(\hat{y_k}) \tag{2}$$

where (x_i, y_i) is either from D_0 or D_1. To be noted, we do not require datasets to contain emotion labels of all cultures to train the emotion classifier E.

For the culture classifier C, we employ another neural network based classifier to predict the culture probility $\hat{t}$:

$$\hat{t} = \text{sigmoid}(\text{C}(z; \theta_C)) \tag{3}$$

To train parameter θ_C in C, the culture classification label $t \in \{0, 1\}$ is self-supervised. For $\mathbf{x}_0 \in D_0$, the label t is 0, and for $\mathbf{x}_1 \in D_1$, $t = 1$, which requires no human annotation. So the culture loss function is:

$$L_{cul} = -\mathbb{E}_{x_1} log(\hat{t}) - \mathbb{E}_{x_0} log(1 - \hat{t}) \tag{4}$$

Since the goal of G is to increase emotion classification and decrease culture classification, the loss function of G is as follows with hyper parameter α to balance the two losses:

$$L_{gen} = L_{emo} - \alpha L_{cul} \quad (5)$$

The adversarial training procedures of the three modules are presented in Algorithm 1. To stabilize the training, we first train the culture classifier C for S_C iterations and then fix C to train E and G for S_G iterations. After the adversarial training, the G is able to generate culture-invariant emotion-salient feature z and E can be used for emotion classification in cross-culture scenarios based on the z.

Algorithm 1 Adversarial Training Procedure

Require: feature generator G, emotion classifier E, and the culture classifier C;
Input: (x_0, y_0) annotated video emotion pair in culture D_0, (x_1, y_1) or x_1 in culture D_1;
for epoch = 0, ..., N **do**
  for iter = 0, ..., S_C **do**
    Compute culture loss L_{cul} using Eq. 4
    Adam update θ_C with L_{cul}
  end for
  for iter = 0, ..., S_G **do**
    Compute L_{emo} and L_{gen} using Eq. 2 and Eq. 5
    Adam update θ_E with L_{emo}
    Adam update θ_G with L_{gen}
  end for
end for

4 Experiments

4.1 Dataset

We utilize two video emotion datasets, CHEAVD 2.0 and AFEW, in Chinese and English cultures respectively.

CHEAVD 2.0 [20] is used in the MEC 2017 challenge, which contains 7,032 video segments collected from Chinese films, TV plays and talk shows. The videos are divided into 3 parts: 4,918 for training, 708 for validation and 1,406 for testing. Each video is annotated with one of the eight possible emotion categories including angry, happy, sad, neutral, surprise, worried, anxious and disgust.

AFEW [21] is used in the EmotiW 2017 challenge. The videos are extracted from English films and TV shows. There are 773 videos for training, 383 for validation and 652 for testing. The annotations are seven possible emotion categories including angry, disgust, fear, happy, neutral, sad and surprise.

We select four common emotions in both CHEAVD 2.0 and AFEW datasets, namely neural, happy, angry and sad. Since the emotion distribution in CHEAVD 2.0 dataset is quite unbalanced, we sample videos in this dataset to match the emotion distribution in AFEW. The statistics of the two datasets are presented in Table 1.

Table 1. The number of video segment of emotion categories on CHEAVD 2.0 and AFEW.

Corpus	Set	Neu	Hap	Ang	Sad	Total
CHEAVD 2.0	Train	600	524	507	300	1931
	Val	240	169	202	105	716
	Test	360	254	303	124	1041
AFEW	Train	104	110	100	100	414
	Val	40	40	40	35	155
	Test	63	63	57	43	226

4.2 Experimental Setup

We evaluate our proposed model in two settings:
1) **cross-culture:** the emotion model is trained on one culture but tested on the other culture. In training, data from both culture is fed into culture classifier but data from only one culture is fed into emotion classifier. In testing, data from the other culture is used for performance evaluation. This setting is used to evaluate the generalizability of the emotion model in a novel culture. For example, the baseline is to directly apply G and E optimized on training data of D_0 to the testing videos of D_1. For our proposed adversarial model, G is not only optimized with emotion labels of D_0, but also is trained to confuse C to remove the culture influence.
2) **multi-culture:** the emotion model is trained with the combination of two cultures such as D_0 and D_1. This setting is used to evaluate the emotion recognition performance when multi-cultural data is available. We use the G and E directly trained on D_0 and D_1 as the baseline. Since the size of AFEW dataset is much smaller than that of CHEAVD 2.0 dataset, we upsample the AFEW dataset during training which can improve the performance in the multi-culture setting.

We set G with two hidden layers and the number of hidden units of each layer is 512 and 256 respectively. The E and C are of the same structure with one hidden layer of 500 hidden units. Dropout is used for each fully connected layer to avoid over-fitting. The α in Eq. 5 is set as 1. For training balance between emotion classifier and culture classifier, the iteration number S_C and S_G is set to 1 and 2 respectively. We apply Adam algorithm [22] with learning rate of 1e-4 to optimize the neural networks. We use accuracy and macro F1 [23] as the evaluation metrics.

4.3 Experimental Results

We first present the emotion recognition performance in the mono-culture setting in Table 2, which utilizes dataset of the same culture for training and testing. It serves as an upper-bound for the cross-culture setting and a lower-bound for the multi-culture setting. As we can see, the Dface feature achieves better emotion classification performance than the IS10 feature. But the two types of features are complementary for video emotion recognition. The early fusion of them can boost the recognition performance significantly.

In Table 3, we present the cross-culture emotion recognition performance. The baseline method simply applies the

Table 2. Emotion recognition performance in mono-cultural setting on the testing set of CHEAVD 2.0 and AFEW.

Test / Feature	CHEAVD 2.0		AFEW	
	ACC	MAF1	ACC	MAF1
IS10	56.33	55.16	52.12	52.04
Dface	64.76	61.59	60.80	60.55
IS10+Dface	**69.95**	**69.11**	**67.88**	**67.77**

Table 3. Emotion recognition performance in cross-culture setting on CHEAVD 2.0 and AFEW testing sets. BSL refers to the baseline and ADV refers to the proposed adversarial approach.

Test / Feature	model	CHEAVD 2.0		AFEW	
		ACC	MAF1	ACC	MAF1
IS10	BSL	41.27	40.96	47.17	46.95
	ADV	42.36	41.24	50.88	49.84
Dface	BSL	51.68	51.82	54.96	52.81
	ADV	53.89	52.64	59.29	58.54
IS10+Dface	BSL	54.89	55.59	63.54	62.83
	ADV	**55.91**	**56.98**	**65.49**	**64.54**

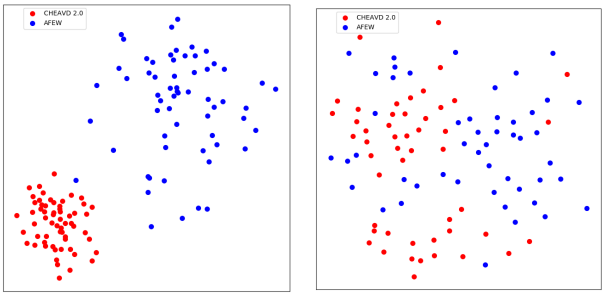
emotion model trained in the other culture for testing. Due to the culture discrepancy, the emotion recognition performance of the baseline model drops dramatically compared with the mono-culture setting in Table 2. Since the size of CHEAVD 2.0 training set is much larger than the AFEW training set, the model transferring from CHEAVD 2.0 to AFEW is better than the inverse direction. Our proposed adversarial model learns to suppress the culture influence on the emotion representation, which leads to significant improvements over the baseline on all modality features. On the AFEW testing set, the performances of the adversarial training model are on par with the model trained in the mono-culture. The results on the cross-culture setting demonstrate the effectiveness of the proposed approach to alleviate culture bias, which can generalize to recognizing emotions in novel cultures.

The performance of multi-culture setting is shown at Table 4. Though we embrace more data with the joint training of two datasets, the emotion recognition performance of the baseline is only on par with the mono-culture setting if not worse. This suggests that the simple multi-culture dataset combination strategy suffers from the discrepancies of different culture datasets for emotion recognition. By using the adversarial training strategy, we achieve accuracy of 70.50% on the Chinese CHEAVD 2.0 dataset and 68.14% on the English AFEW dataset, which significantly outperforms the baseline performance. It demonstrates that the adversarial model can narrow the culture gap between different cultures for emotion recognition and take the advantage of multi-culture data to improve performance of each culture.

We analyze the emotion representation learned from G in the cross-culture setting in Figure2 with t-SNE. In the baseline method, the features in Chinese and English cultures are in different domains which results in poor cross-culture performance. However, by the adversarial learning, the features from the two features are mixed, which demonstrates that our adversarial learning can project multimodal features into similar emotion space and narrow the gap of different cultures.

Table 4. Emotion recognition performance in multi-culture setting on CHEAVD 2.0 and AFEW testing sets. BSL refers to the baseline and ADV refers to the proposed adversarial approach.

Test / Feature	model	CHEAVD 2.0		AFEW	
		ACC	MAF1	ACC	MAF1
IS10	BSL	52.39	50.83	51.24	50.58
	ADV	54.65	53.53	53.32	52.88
Dface	BSL	62.57	59.89	60.00	59.04
	ADV	65.03	62.62	61.50	60.83
IS10+Dface	BSL	69.74	68.91	67.88	67.76
	ADV	**70.50**	**69.69**	**68.14**	**68.45**



(a) Cross-BSL (b) Cross-ADV

Fig. 2. Visualization of emotion representations from G in the cross-cultural setting.

5 Conclusion

We present a novel adversarial framework to improve the generalization of multimodal emotion models in multi-culture scenario. Through the adversarial competition of the emotion recognition and culture recognition, the learned features are more emotion salient to improve the emotion recognition performance and culture invariant to confuse the culture classification. The proposed adversarial approach outperforms the baselines in both cross-culture and multi-culture evaluation settings. In the future, we will explore the adversarial approach to suppress other irrelevant factors such as gender and personality to further improve emotion recognition.

6 Acknowledgments

This work was supported by National Key Research and Development Plan under Grant No. 2016YFB1001202, Research Foundation of Beijing Municipal Science & Technology Commission under Grant No. Z181100008918002 and National Natural Science Foundation of China under Grant No. 61772535.

7 References

[1] Morena Danieli, Giuseppe Riccardi, and Firoj Alam, "Emotion unfolding and affective scenes:a case study in spoken conversations," in *Icmi-workshop on Emotion Representations Modelling for Companion Technologies*, 2015, pp. 5–11.

[2] Richard J Gerrig, Philip G Zimbardo, Philip Georg Zimbardo, Etats-Unis Psychologue, and Philip Georg Zimbardo, *Psychology and life*, Pearson, 2010.

[3] Silvia Monica Feraru, Dagmar Schuller, and Björn Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 125–131.

[4] Michael Neumann and Ngoc Thang Vu, "Cross-lingual and multilingual speech emotion recognition on english and french," *ICASSP*, 2018.

[5] Bo-Chang Chiou and Chia-Ping Chen, "Speech emotion recognition with cross-lingual databases," in *INTERSPEECH 2014*, 2014, pp. 558–561.

[6] Shizhe Chen, Shuai Wang, and Qin Jin, "Multimodal emotion recognition in multi-cultural conditions," in *Journal of Software*, No. 4, 2018, pp. 1060–1070.

[7] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *International Workshop on Audio/visual Emotion Challenge*, 2016, pp. 97–104.

[8] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shiri Saleem, Rohi Kumar, Vembu Aravind Namandi, and Prasad Rohit, "Emotion recognition using acoustic and lexical feature," in *INTERSPEECH 2012*, 2012, pp. 366–369.

[9] Shizhe Chen and Qin Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," *ACM Multimedia*, 2016.

[10] Pavel Matejka Hesam Sagha, "Enhancing multilingual recognition of emotion in speech by language identification," in *INTERSPEECH 2016*, 2016.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[12] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," 2017.

[13] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[14] Florian Eyben, "Opensmile:the munich versatile and fast open-source audio feature extractor," in *ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[15] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[16] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[18] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[19] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *The Workshop on Audio/visual Emotion Challenge*, 2017, pp. 19–26.

[20] Ya Li, Jianhua Tao, Björn Schuller, Shiguang Shan, Dongmei Jiang, and Jia Jia, "Mec 2017: Multimodal emotion recognition challenge," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–5.

[21] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al., "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[23] Yutaka Sasaki et al., "The truth of the f-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.