Towards Intercultural Affect Recognition: Audio-Visual Affect Recognition in the Wild Across Six Cultures

Leena Mathur[1], Ralph Adolphs[2], and Maja J Matarić[1]
[1] University of Southern California, [2] California Institute of Technology

***Abstract*— In our multicultural world, affect-aware AI systems that support humans need the ability to perceive affect across variations in emotion expression patterns across cultures. These systems must perform well in cultural contexts without annotated affect datasets available for training models. A standard assumption in affective computing is that affect recognition models trained and used within the same culture (*intracultural*) will perform better than models trained on one culture and used on different cultures (*intercultural*). We test this assumption and present the first systematic study of intercultural affect recognition models using videos of real-world dyadic interactions from six cultures. We develop an attention-based feature selection approach under temporal causal discovery to identify behavioral cues that can be leveraged in intercultural affect recognition models. Across all six cultures, our findings demonstrate that intercultural affect recognition models were as effective or more effective than intracultural models. We identify and contribute useful behavioral features for intercultural affect recognition; facial features from the visual modality were more useful than the audio modality in this study's context. Our paper presents a proof-of-concept and motivation for the future development of intercultural affect recognition systems, especially those deployed in low-resource situations without annotated data.**

I. INTRODUCTION

Advances in affective computing, multimodal machine learning, and social signal processing are enabling the development of automated systems that can sense, perceive, and respond to human affective states [3], [33], [51]. *Affect* refers to neurophysiological states that can function as components of emotions or longer-term moods. Psychologists and neuroscientists view affective states as latent variables that must be inferred from measured variables across communication modalities, such as facial expressions, body postures, and tone of voice [1], [5]. Affect is typically represented along two dimensions: how pleasant or unpleasant each state is (valence) and how passive or active each state is (arousal) [38]. AI systems that are *affect-aware*, possessing the ability to estimate human affect, can enhance the ability of robots and virtual agents to support human health, education, and well-being [8], [20], [37].

While affect-aware AI systems have the potential to help humans, current affect recognition approaches do not perform well across different cultures; inaccuracies are an issue, particularly, for cultures with limited data available for training affect recognition models. This key, underexplored challenge in affective computing is thought to arise due to differences across cultures in behavioral cues and expression norms that indicate affective states (e.g., facial movements, voice tone) [9], [22], [28]. Collecting large amounts of video data across cultures and annotating this data for affect can be expensive, time-consuming, and, for underrepresented cultures, infeasible. This motivates the need for affect recognition approaches that can adapt and perform well across cultures without annotated affect data available.

We address this challenge by developing audio-visual affect recognition models with *attention-based feature selection* (ABFS) under temporal causal discovery [31]. Our approach can be viewed as feature-based unsupervised domain adaptation [25] that identifies potential causal feature relationships to be used by models for affect recognition in cultures on which they are not trained. Our approach is motivated by the idea that potential causal relationships between affect and behavioral cues (facial and vocal) in one cultural context may be robust to spurious culture-specific noise, allowing an affect recognition model with ABFS to perform well in cultures on which it is not trained.

We conducted experiments with SEWA, the largest publicly-available multicultural affect video dataset [24], that contains real-world dyadic interactions of participants from 6 cultures: British, Chinese, German, Greek, Hungarian, and Serbian. This paper presents the first systematic study of *intercultural* affect recognition models that are trained on videos from one culture and tested on videos from different cultures. We compare the performance of these *intercultural* models to *intracultural* affect recognition models that are trained and tested on videos from the same culture.

Given the existence of culture-specific emotion expression patterns [9], [22], [28], it might be expected that *intracultural* affect recognition models will perform better than *intercultural* models. We test this assumption in our research. Our findings across all six cultures, surprisingly, suggest that intercultural affect recognition models may be as effective and, in some domains, more effective than intracultural models. Our results contribute new baseline findings and a proof-of-concept for the potential of creating *intercultural affect recognition* systems that can be used across cultures. This work makes the following contributions:

- The first systematic study of audio-visual intercultural affect recognition models, contributing a new baseline.
- New findings regarding the potential of intercultural affect recognition models to match or outperform intracultural models.
- Analysis and identification of automatically-selected interpretable features, particularly facial cues, that were useful for affect recognition across cultural domains.

II. RELATED WORKS

A. Culture and Affect Expression

Humans within a *culture* typically share common beliefs, values, and social norms that can influence their perception of the world and communication patterns [7]. Scientists since the 1800s [10] have debated the extent to which processes for affect expression are *culturally-universal* versus *culture-specific*. A leading historical perspective is that affect expression patterns tend to be culturally-universal. Studies from participants in Argentina, Brazil, Chile, Japan, New Guinea, and the United States [12]–[16] identified Facial Action Unit (FAU) configurations that were commonly observed across cultures when participants conveyed basic discrete emotions (e.g., happy, sad). However, these studies did not consider the existence of culture-specific display rules that influence whether it is appropriate, in a given cultural context, to display affective information. Subsequent studies have not supported the view that affect expression patterns are culturally-universal [19]. Psychology studies have found that affect expression patterns in face and eye movements vary within cultures and vary even more across cultures [9], [22], [40]. Automated analysis of affect expression from images in the wild found very few facial affect expression patterns shared across cultures [46]. Since affect expression is influenced by cognitive appraisal mechanisms [2], [27], researchers have posited that differences in cultural value systems and norms may result in culture-specific cognitive appraisal mechanisms for expressing affect [39].

B. Continuous Affect Recognition Models Across Cultures

Given the culture-specific aspects of affect expression patterns, researchers in affective computing have started to acknowledge the importance of creating affect recognition models that can perform well across cultures; this area remains an open research problem [44]. A seminal data collection effort to address this problem occurred through the creation of the SEWA database, the first and largest publicly-available video dataset of affect in the wild, including six cultures [24]. We use SEWA data in this paper.

Prior work using a subset of the SEWA database trained audio-visual affect recognition Long Short-Term Memory (LSTM) networks on videos of German participants to predict valence and arousal in videos of Hungarian participants [34]. Subsequent research efforts leveraged semi-supervised learning on a larger subset of the SEWA database to train recurrent models on videos from German and Hungarian participants and test these models on Chinese participants [29]. Another research effort [35] trained shallow LSTM networks on videos from German and Hungarian participants and tested on Chinese participants; this study found that facial features, specifically FAUs, were useful for predicting valence and arousal. These findings motivated us to analyze the contributions of FAU in our paper's experiments. Prior researchers have also investigated elastic weight consolidation for intercultural affect recognition across French and German cultures [21], with German data from a subset of the SEWA database and French data from the RECOLA database [36]. These prior works did not use data from all six cultures and did not explore methods for selecting useful features for affect recognition across cultures. Our paper, to the best of our knowledge, contributes the first study of intercultural affect recognition models across all six cultures in SEWA, as well as the first study using ABFS approaches to identify effective behavioral cues for intercultural, audio-visual affect recognition.

C. Causal Discovery in Time-Series Data

Continuous affect recognition is typically viewed as a multivariate time-series prediction problem, with the goal of predicting affect labels at each timestep of a video. Models that rely on causal feature representations have the potential to capture the underlying dynamics of a domain and to be robust to spurious noise, making them useful for tasks that involve transferring knowledge across domains [42] (e.g., intercultural affect recognition). Various approaches exist for inferring causal features from time-series samples, including constraint-based methods that estimate a single causal graph through conditional independence testing [45], score-based methods that optimize a chosen objective when learning a causal graph [6], and causal graph estimation methods that use gradient-based learning to estimate a causal graph [26], [31], [53]. We adapt an attention-based causal graph estimation method [31] to identify features that can be useful for affect recognition across cultures. Our approach is motivated by the idea that potential causal relationships between affect and behavioral cues in one culture may be robust to spurious culture-specific noise, supporting the feasibility of intercultural affect recognition models. To the best of our knowledge, we contribute the first study of causal graph estimation approaches for identifying useful features in continuous intercultural affect recognition.

III. METHODOLOGY

We conducted a systematic study of intercultural video-based affect recognition with ABFS; an overview of the modeling approach is visualized in **Fig. 1**.

A. Multicultural Affect Video Dataset

We used the SEWA Database [24], the largest video dataset for in-the-wild affect estimation. SEWA contains videos of individuals engaged in naturalistic, real-world dyadic interactions. Aligned with prior studies [23], [24], [48], we chose to use the Basic SEWA database (538 video clips, 275 individuals), which contains the most comprehensively annotated subset of videos within SEWA. The individuals ranging from 18-40 years, have a balanced gender representation (52% male, 48% female), and come from 6 cultures: British, Chinese, German, Greek, Hungarian, and Serbian. The number of participants varied across cultures within the range of 35 to 56. The valence and arousal of each individual has been labeled at each frame in the videos; the provided annotations are normalized in the continuous range [0, 1]. The final affect labels are an aligned

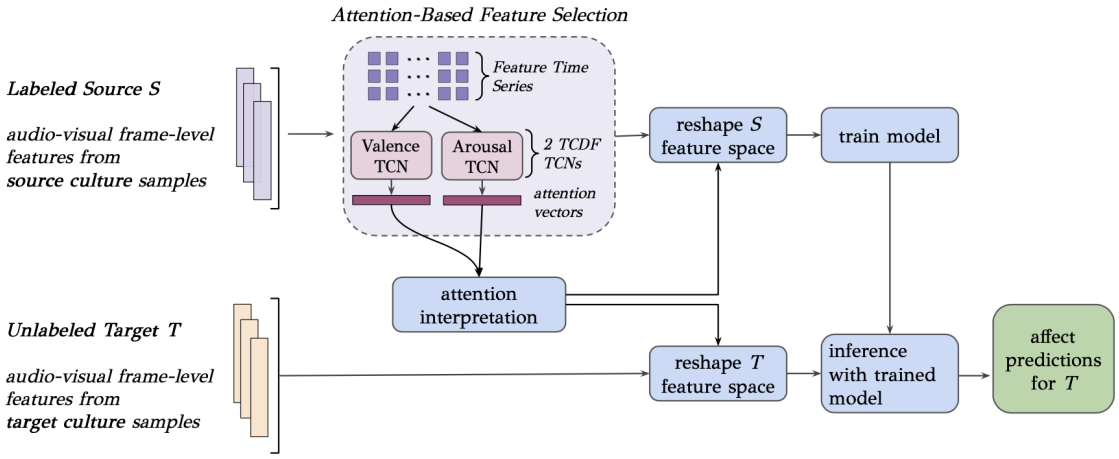

Fig. 1. Visualization of affect recognition models across culture domains with Attention-Based Feature Selection (ABFS).

aggregation (canonical time warping [50]) of the frame-level annotations from 5 annotators who shared the same culture as the person they were labeling, boosting the validity of the affect labels. It is worth noting that the ground truth on which models are trained is an aggregation of the annotators' perceptions of the affect of people in videos; this method for obtaining ground truth is the current norm in affective computing [11]. Within the Basic SEWA data, we computed the correlation (threshold = 0) of the affect annotations across annotators to identify a final subset of 510 videos.

B. Multimodal Feature Extraction

We extracted 792 audio-visual features from the speakers in each video frame to represent each speaker's observable behavioral cues over time. The OpenSMILE toolkit (version 2.0) [18] was used to extract 83 audio features that capture cepstral, spectral, prosodic, energy, and voice quality information from raw speech signals at each audio frame (10 ms step size). 18 audio features came from the GeMAPS feature set, and 65 features came from the ComParE feature set; prior research found these signals to be useful at capturing affective properties of speech [17], [24], [43], motivating their use in our work. The OpenFace toolkit14 (version 2.2.0) [4] was used to extract 709 visual features that capture eye gaze, FAUs, and head pose information from speakers from each video frame. After obtaining these audio-visual features, we aligned audio features, visual features, and affect labels at each video frame. Similar to prior research [47], we focused on using interpretable audio-visual features from OpenSMILE and OpenFace, instead of using deep feature representations, to more effectively identify and analyze behavioral cues in our experiments.

C. Problem Formulation for Intercultural Affect Recognition

We approach intercultural affect recognition as a feature-based unsupervised domain adaptation (UDA) problem [25], [52]. This type of UDA involves training models on labeled source domains and re-shaping the feature space in order to perform tasks on unlabeled target domains. In this paper, a domain refers to a set of videos from a single culture.

Given a set of source sample videos S with affect labels at each frame and a set of unlabeled target sample videos T, the goal is to train a model on S that performs well at predicting affect at each frame of T with minimal error when S and T come from different cultures. Each sample in S and T is padded to the same length of $L = 1000$ and has the same initial number of features $D = 792$. Each S contains N time-series representations of D-dimensional videos, designated as $SV_1 ... SV_N$, where each $SV_{1...N} \in R^{LXD}$ and each SV has corresponding affect time-series labels for valence and arousal, continuous between [0, 1]. Each T contains M unlabeled video time-series representations designated as $TV_1 ... TV_M$, where each $TV_{1...M} \in R^{LXD}$. We use ABFS (described in **Section III-D**) on samples in S to identify potential causal relationships between the features in S and labels in S. We then reshape the feature space of both S and T to include these features. After training our final model on the re-shaped samples of S, we perform inference with the trained model on the re-shaped samples of T to obtain affect predictions for the unlabeled T samples.

D. Attention-Based Feature Selection

We develop an ABFS approach under the temporal causal discovery framework (TCDF) [31] to identify potential causal relationships in each source culture between the audio-visual time-series features and valence and arousal affect labels. Across source samples, we train two separate temporal convolutional neural networks (TCNs), the Valence TCN and the Arousal TCN, to predict valence and arousal at each video frame by using only the past values of the audio-visual features until that frame. Similar to the original paper [31] we fix TCN hyperparameters to make network initialization constant across all experiments, reported for reproducibility (epochs=1000, kernel size=250, mean square error loss, adam optimizer, dilation coefficient=250).

Each TCN includes a separate channel for each audio-visual time series and includes an attention mechanism that computes attention scores for each audio-visual feature. The Valence TCN and Arousal TCN have trainable attention

vectors that are element-wise multiplied by each audio-visual time series feature. Therefore, the final attention vectors for each TCN capture how much attention that TCN pays to each audio-visual feature when predicting valence or arousal. Similar to [31], we apply a discrete threshold (of 0.25) on these final attention scores to identify potentially causal features. We refer to these selected ABFS features as *potentially causal* so that our research does not make ungrounded assumptions about causality in the studied affect context. Further details regarding the TCDF framework are found in [31]. We combine ABFS features selected from the Valence TCN and Arousal TCN into the final feature set used to reshape the source and target domains in intercultural affect recognition, as described in **Section III-C**.

E. Model Training and Metrics

Using the 6 cultures present in the SEWA database, we conduct 30 experiments with *intercultural* affect recognition models that are trained on videos from one culture and tested on videos from a different culture. Our *intercultural* affect recognition models are baselined against 6 *intracultural* models that are trained and tested on videos from the same culture. For a point of comparison, similar to [24], we trained and tested a *multicultural* model on groups of videos that contained a random mixture of all cultures.

All 37 experiments were conducted with 5-fold cross-validation. Within each cross-validation experiment, we standardized the features in both the training and testing set by the distributions in the training set. Aligned with prior SEWA research [35], all models were LSTM networks [41], chosen for their potential to capture nonlinear dependencies in time-series data. Similar to prior SEWA research [35], we fix LSTM hyperparameters to make network initialization constant across all 37 experiments, reported for reproduciblity (dropout = 0.1, learning rate = 0.1, hidden units = 256, adam optimizer, epochs = 1000 with early stopping). Models were implemented in PyTorch [32].

For each experiment, two metrics were computed at each cross-validation fold to evaluate the model's affect predictions with respect to ground truth: (1) RMSE, the Root mean square error minimized between predictions and ground truth; (2) CCC, the concordance correlation coefficient. Similar to prior SEWA research [49], we trained models with RMSE and CCC loss functions. We used RMSE to compare the prediction performances of the models.

IV. RESULTS AND DISCUSSION

Results from the affect recognition experiments are presented in **Table I**. Our analysis focuses on the RMSE metric. A comparison of intercultural affect modeling results (RMSE) relative to intracultural results for each target culture are visualized in **Fig. 2**. For 5 of the 6 target cultures, British, Chinese, German, Greek, and Hungarian, each intercultural affect recognition model outperformed or matched the corresponding intracultural model. In the Serbian case, the intercultural models closely matched the performance of the intracultural model. We found that 27/30 intercultural models were equivalent to or outperformed their intracultural models, and 19/30 intercultural models outperformed the multicultural model. *Our findings suggest that intercultural affect modeling may be as effective and, in some domains, more effective than intracultural modeling.*

The intercultural affect recognition approach with ABFS appears to have selected features within each culture that were less-influenced by culture-specific noise (e.g., display rules [30]). To identify the key behavioral cues that enabled intercultural affect recognition models to transfer knowledge across cultures and outperform intracultural models, we examined the ABFS features selected by each model. **Table II** lists these features, along with feature descriptions. Across all source cultures, all of the selected ABFS features were from different feature sets within the visual modality. It appears the visual modality was more useful than the audio modality for affect recognition in the studied SEWA context. Within the 15 selected features from the visual modality, there were 4 features from PDM parameters (face shape representations), 2 features from head pose movements, 4 features from eye and eyebrow movements, and 5 features from lip, nose, and cheek movements. These findings on feature importance align with a prior study [35] that found FAU features important for affect recognition with a smaller subset of the SEWA dataset. Our findings indicate that temporal patterns in facial movements have potential for use in intercultural affect recognition models.

TABLE I
RESULTS (RMSE/CCC) FROM INTRACULTURAL, INTERCULTURAL, AND MULTICULTURAL AFFECT RECOGNITION MODELS. (INTRACULTURAL RESULTS ARE IN YELLOW ON THE DIAGONAL; GREEN HIGHLIGHTS INDICATE THAT THE INTERCULTURAL MODEL WAS EQUIVALENT TO OR OUTPERFORMED THE CORRESPONDING INTRACULTURAL MODEL; MULTICULTURAL RESULTS ARE IN BLUE).

		Target Culture						
		British	**Chinese**	**German**	**Greek**	**Hungarian**	**Serbian**	**All**
Source Culture	**British**	0.39 / 0.06	0.23 / -0.06	0.23 / 0.05	0.27 / 0.04	0.38 / -0.03	0.21 / -0.01	
	Chinese	0.36 / -0.03	0.25 / 0.02	0.24 / -0.03	0.27 / -0.02	0.37 / 0.01	0.21 / 0.00	
	German	0.35 / 0.01	0.22 / -0.01	0.25 / 0.01	0.27 / 0.01	0.37 / 0.00	0.20 / 0.00	
	Greek	0.35 / 0.02	0.21 / -0.03	0.23 / 0.02	0.31 / 0.02	0.37 / -0.01	0.20 / 0.00	
	Hungarian	0.36 / -0.03	0.25 / 0.02	0.24 / -0.02	0.30 / -0.03	0.51 / 0.05	0.21 / 0.00	
	Serbian	0.35 / 0.00	0.22 / 0.00	0.22 / 0.00	0.27 / 0.00	0.37 / 0.00	0.20 / 0.00	
	All							0.30 / 0.00

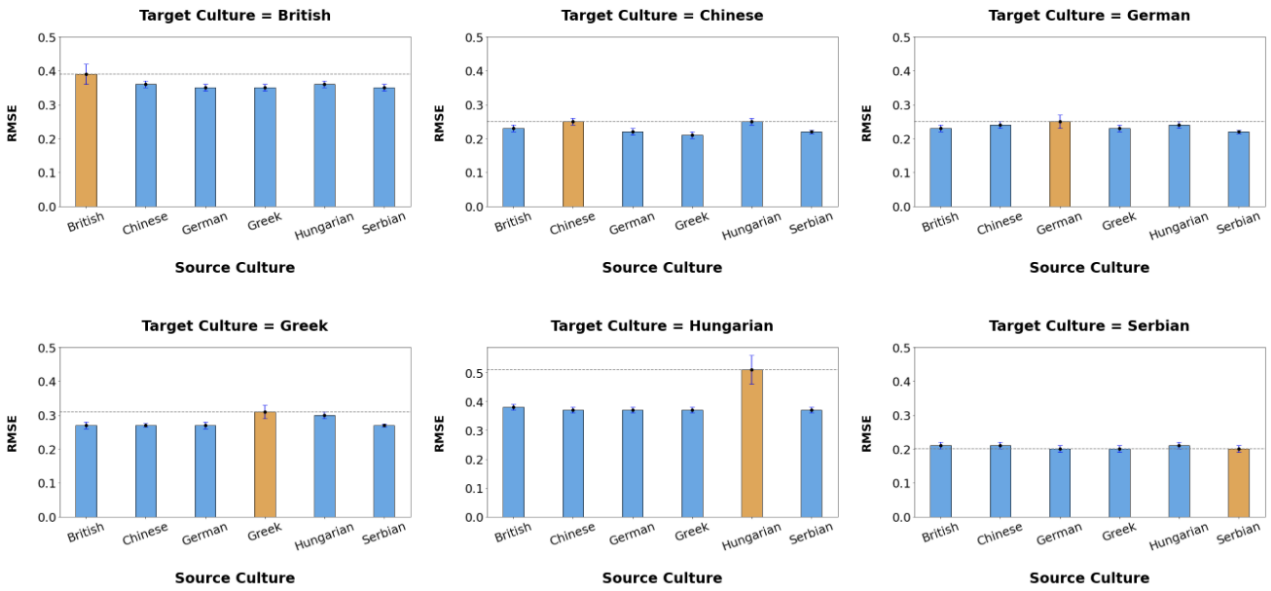

Fig. 2. Comparison of intercultural affect recognition performance (RMSE) with intracultural affect recognition performance (intracultural results are in orange, intercultural results are in blue, error bars are displayed)

TABLE II
FEATURES SELECTED BY THE ABFS MODELS.

Source	ABFS Features Selected and Used
British	PDM parameter 10 (deformation due to expression)
Chinese	Head pose roll in the Z axis (pose Rz)
	Left eye gaze direction, x coordinate (Gaze 0 X)
German	Intensity lip tightener (FAU 23)
	Left lip landmark (X 60)
	Left cheek landmark (X 1)
Greek	PDM parameters 2 (deformation due to expression)
	PDM parameter 27 (deformation due to expression)
	Intensity of inner brow raise (FAU 1)
	Head pose roll along the Z axis (pose Rz)
Hungarian	nose landmark (X 30, moving with head movement)
	Lip corner pull intensity (FAU 12)
	Eye gaze direction for right eye (Gaze 1 X)
Serbian	PDM parameter 16 (deformation due to expression)
	Right eye landmark (X 44)

Our analysis focused on RMSE, a standard metric for time-series prediction, to compare the prediction performance of intercultural and intracultural affect recognition models. We report CCC for each experiment for consistency with prior papers that used subsets of SEWA and included CCC [24], [34], [35]. We contribute new RMSE and CCC findings for intercultural affect recognition across all six cultures in the SEWA data. Prior works focused on only two or three of the six cultures, each leveraging different training, validation, and testing splits. These differences constrained our ability to make direct comparisons with prior papers. Low CCC in our models motivates future work to create affect recognition techniques that perform well in minimizing regression error and maximizing correlation metrics across cultures.

V. CONCLUSION

Our paper presents the first systematic study of audio-visual intercultural affect recognition models, using videos of real-world dyadic interactions from six cultures. We contribute new findings regarding the potential for intercultural affect recognition models to match or outperform intracultural models. Our results serve as a baseline, proof-of-concept, and motivation for the future development of intercultural affect recognition systems that can be deployed in cultures without annotated affect datasets for training. Through our ABFS approach, we identified and analyzed the automatically-selected features (primarily facial cues) that were useful for affect recognition across cultures; these findings inform future approaches for video-based intercultural affect recognition.

To support the robustness and generalizability of intercultural affect recognition findings, future work could include collecting and experimenting with larger multicultural affect datasets to encompass participants, culture domains, and annotators beyond the six cultures in SEWA. Future work might also venture beyond valence and arousal to explore models for jointly recognizing higher-level affective states (e.g., ”happy”) in the design of intercultural human-machine interaction systems. Our paper informs and motivates the future development of affect-aware AI systems to support humans in the wild across cultures.

VI. ACKNOWLEDGMENTS

This work was supported by a Caltech Summer Undergraduate Research Fellowship. The research in this paper uses the SEWA Database collected in the scope of SEWA project financially supported by the European Community’s Horizon 2020 Programme (H2020/20142020) under Grant agreement No. 645094.

REFERENCES

[1] R. Adolphs and D. Andler. Investigating emotions as functional states distinct from feelings. *Emotion Rev.*, 10(3):191–201, 2018.

[2] M. B. Arnold. Emotion and personality. 1960.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 41(2):423–443, 2018.

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior anal. toolkit. In *IEEE Int. Conf. on Autom. Face & Gesture Recognit.*, pages 59–66. IEEE, 2018.

[5] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Science in the Public Interest*, 20(1):1–68, 2019.

[6] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, et al. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[7] H. Betancourt and S. R. López. The study of culture, ethnicity, and race in amer. psychol. *Amer. Psychologist*, 48(6):629, 1993.

[8] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas. *The Oxford Handbook of Affective Computing*. Oxford Library of Psychol., 2015.

[9] L. A. Camras, R. Bakeman, Y. Chen, K. Norris, and T. R. Cain. Culture, ethnicity, and children's facial expressions: a study of european amer., mainland chinese, chinese amer., and adopted chinese girls. *Emotion*, 6(1):103, 2006.

[10] C. Darwin and P. Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[11] S. D'Mello, A. Kappas, and J. Gratch. The affective computing approach to affect measurement. *Emotion Rev.*, 10(2):174–183, 2018.

[12] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symp. on Motivation*, 1971.

[13] P. Ekman. Cross-cultural studies of facial expression. *Darwin and facial expression: A century of res.in rev.*, 169222(1), 1973.

[14] P. Ekman. Facial expression and emotion. *Amer. Psychologist*, 48(4):384, 1993.

[15] P. Ekman, W. V. Friesen, and S. S. Tomkins. Facial affect scoring technique: A first validity study. 1971.

[16] P. Ekman, E. R. Sorenson, and W. V. Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice res.and affective computing. *IEEE Trans. on Affect. Comput.*, 7(2):190–202, 2015.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM Int. Conf. on Multimedia*, pages 835–838, 2013.

[19] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251, 2014.

[20] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, et al. Affect.personalization of a social robot tutor for children's second language skills. In *AAAI Conf. on Artif. Intell.*, volume 30, 2016.

[21] J. Han, Z. Zhang, M. Pantic, and B. Schuller. Internet of emotional people: Towards continual affect. computing cross cultures via audiovisual signals. *Future Gener. Comp. Syst.*, 114:294–306, 2021.

[22] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns. Facial expressions of emotion are not culturally universal. *PNAS*, 109(19):7241–7244, 2012.

[23] J. Kossaifi, A. Toisoul, A. Bulat, Y. Panagakis, T. M. Hospedales, and M. Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *IEEE/CVF Conf. on Comp. Vision and Pattern Recognit.*, pages 6060–6069, 2020.

[24] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, et al. Sewa db: A rich database for audio-visual emotion and sentiment res.in the wild. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 43(3):1022–1040, 2019.

[25] W. M. Kouw and M. Loog. A rev. of domain adaptation without target labels. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 43(3):766–785, 2019.

[26] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[27] R. S. Lazarus. Emotions and adaptation: Conceptual and empirical relations. In *Nebraska Symp. on Motivation*, 1968.

[28] N. Lim. Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integr. Medicine Res.*, 5(2):105–109, 2016.

[29] A. Mallol-Ragolta, N. Cummins, and B. W. Schuller. An investigation of cross-cultural semi-supervised learning for continuous affect recognition. In *INTERSPEECH*, pages 511–515, 2020.

[30] D. Matsumoto. Cultural similarities and differences in display rules. *Motivation and Emotion*, 14(3):195–214, 1990.

[31] M. Nauta, D. Bucur, and C. Seifert. Causal discovery with attention-based convolutional neural networks. *Mach. Learning and Knowl. Extraction*, 1(1):19, 2019.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, et al. Automatic differentiation in pytorch. 2017.

[33] R. W. Picard. *Affective computing*. MIT press, 2000.

[34] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proc. of the ACM MM AVEC Challenge and Workshop*, pages 3–13, 2018.

[35] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proc. of the ACM MM AVEC Challenge and Workshop*, pages 3–12, 2019.

[36] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE Int. Conf. and Workshops on Autom. Face and Gesture Recognit.*, pages 1–8. IEEE, 2013.

[37] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robot.*, 3(19), 2018.

[38] J. A. Russell. A circumplex model of affect. *J. of Person. and Social Psychol.*, 39(6):1161, 1980.

[39] K. R. Scherer and T. Brosch. Culture-specific appraisal biases contribute to emotion dispositions. *Eur. J. of Person.*, 23(3):265–288, 2009.

[40] K. R. Scherer and H. Ellgring. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7(1):113, 2007.

[41] J. Schmidhuber, S. Hochreiter, et al. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[42] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, et al. Toward causal representation learning. *Proc. of the IEEE*, 109(5):612–634, 2021.

[43] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, et al. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *INTERSPEECH*, 2013.

[44] B. W. Schuller. Editorial: Ieee trans. on affect. comput. —challenges and chances. *IEEE Trans. on Aff Comput.*, 8(1):1–2, 2017.

[45] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[46] R. Srinivasan and A. M. Martinez. Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Trans. on Affect. Comput.*, 12(3):707–721, 2018.

[47] L. Tavabi, K. Stefanov, S. Nasihati Gilani, D. Traum, and M. Soleymani. Multimodal learning for identifying opportunities for empathetic responses. In *ACM Int. Conf. on Multimodal Interact.*, pages 95–104, 2019.

[48] M. K. Tellamekala, T. Giesbrecht, and M. Valstar. Modelling stochastic context of audio-visual expressive behaviour with affect. processes. *IEEE Trans. on Affect. Comput.*, 2022.

[49] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Mach. Intell.*, 3(1):42–50, 2021.

[50] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller. Deep canonical time warping. In *IEEE Conf. on Comp. Vision and Pattern Recognit.*, pages 5110–5118, 2016.

[51] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Comput.*, 27(12):1743–1759, 2009.

[52] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. on Intell. Syst. and Technol.*, 11(5):1–46, 2020.

[53] T. Wu, T. Breuel, M. Skuhersky, and J. Kautz. Discovering nonlinear relations with minimum predictive information regularization. *arXiv preprint arXiv:2001.01885*, 2020.