Recognizing Emotions from Out-of-domain Facial Expressions Produced by Non-Actors

Taehyeong Kim Seung Hee Yang Hyunwoong Ko Sungjae Cho
Jun-Young Lee Byoung-Tak Zhang
Seoul National University, Republic of Korea
thkim@bi.snu.ac.kr sy2358@snu.ac.kr powerzines@snu.ac.kr sj.cho@snu.ac.kr
benji@snu.ac.kr btzhang@bi.snu.ac.kr

Abstract

Most of the previous findings on facial emotion recognition are trained and tested using expressions produced by Caucasian adult actors and actresses. The present study reveals that the existing state-of-the-art models underperform when tested on non-actor test datasets, decreasing the performance from 70.57% to 29.05%. We extend the previous work by collecting and training a new model using a wider diversity of training data in terms of the subjects' acting ability, nationality, and age groups, which achieves 46.86% accuracy. The result outperforms human subjects' classification ability by 7.26%. Out-of-domain degradations will be further compromised by applying adaptation techniques when combined with multimodal datasets.

1. Introduction

Facial emotion recognition research has made remarkable progresses in recent years [8, 15, 16, 17] and is one of the most advanced development in applications of artificial intelligence using neural networks. However, many state-of-the-art models are trained and tested on ideal datasets of images expressed by professional actors and actresses. Since they are trained to produce distinct emotional states, the data is well-suited for the classification algorithms.

An overview of often used databases shows that the majority of the existing samples are collected from Caucasian actors and actresses. Although these databases have proven to be valuable resources to quantify emotion, facial stimuli tend to be restricted in ethnicity and age range, such as FERET, MIT, and the Yale databases. While the recent large-scale databases include Asian populations, its facial affects are limited to smile, frown and surprised [6].

There are three types of challenges the state-of-the-art models need to address. First, while most of the real-world facial emotion recognition tasks deal with expressions of non-actors, the trained models use actor-produced images, which may suffer from various types of degradations when tested with emotions produced by non-actors. Second, a survey of the popular facial emotion database [3, 4, 9, 10, 11, 19, 20] show that Caucasians cover the majority of the database, and these models may underperform in samples from other ethnic groups, due to the physiological and cultural differences. Third, children and elderly are under-represented in existing databases. Since age and ethnic factors influence facial expressions, such real-world examples may have negative effects on emotion recognition. Although there exists such expectable gap between the real-world tasks and training examples, to the best of our knowledge, there has been no research efforts towards investigating possible methods that address these challenges.

Figure 1: Examples of video frames. To protect privacy of the subjects, face images are blurred and eye regions are occluded.

The purpose of the present study is to describe the development of a facial expression video database produced by Korean non-actors of various age groups including elderly, and apply the state-of-the-art facial expression recognition technique to this non-ideal dataset.

2. Data Collection and Emotion Recognition

A survey of existing databases highlight the skewedness towards adult Caucasian adult actors' emotions, and there is a need for establishing a non-actor Asian face database. We collected video recordings of six 'basic' emotions, and neutral condition [3, 12]. The current study included clinically normal 174 participants, including 88 females and 86 males

Name	Average accuracy (%)	Training data	Test data
VGG-FER-FER	70.57	FER2013	FER2013
VGG-FER-NAE	29.05	FER2013	NAE
VGG-NAE-NAE	46.86	NAE	NAE
VGG-NAE-FER	25.30	NAE	FER2013
Human-None-NAE	39.60	None	NAE

Table 1: Classification accuracies of models and human subjects in this study. The naming scheme follows the acronyms of "Model-TrainData-TestData" format, where NAE refers to "Non-Actor Emotions"

of ages ranging between 18-84 years. The participants are asked to produce and hold a facial expression for about ten seconds, which is recorded at 30 frames per second. Example images of this database are shown in Figure 1.

We extracted a total of 32,970 images from the videos. Augmentation technique was implented, which alleviates the impacts of misalignment and pose variance, as well as overfitting. Thirty images per video are extracted, resized to 48×48 pixels, randomly rotated by angles between -5° and 5°, and augmented with 44×44 random cropping, which are accompanied by a horizontal flip at the training time.

Convolutional neural networks is known to be very effective for learning face representations given a large number of samples, achieving state-of-the-art results in Facial Expression Recognition Challenge [1]. The network uses facial features to determine the the emotion. The trained model architectures follow VGG19 [18], which starts with an input layer of 44×44, followed by convolutional layers with a batch normalization layer and ReLU unit and max-pooling layers. The network is finished with one fully connected layer, connected to a soft-max output layer.

3. Results and Discussions

The state-of-the-art model [17] achieves 70.57% when trained and tested with facial emotions produced by adult Caucasian actors. However, when the same model is tested on our collected non-ideal database, the average classification accuracy decreases to 29.05%. We confirm our prediction that the models trained on actor data do not perform well on non-actor emotion test sets, and vice versa. On the same test set, the newly trained model with the proposed method achieves 46.86%, as summarized in Table 1 and Figure 2. The degradations observed in non-ideal examples were compensated by retraining on non-ideal data.

We further compare the accuracy of the newly trained model with that of the human subjects. Four human subjects were presented the same held-out images and emotion categories, and matched each face with the word that best describe the emotion. Average accuracy by the human subjects was 39.60%, and VGG-NAE-NAE model outperforms

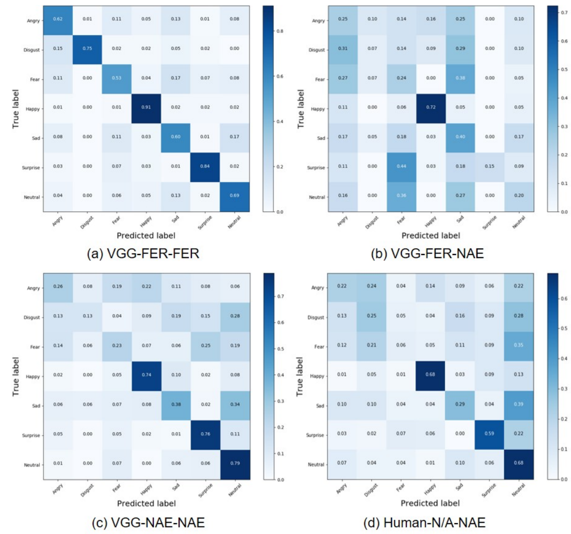

Figure 2: Confusion matrices of recognition models and human subjects.

by 7.26%. The results verify that it is difficult to recognize emotions from non-ideal facial data. Although this maybe a trivial conclusion in domain adaptation problems in neural network trainings, the contribution point is that this study quantitatively verified the issue in an emotion recognition task, during which we collected non-ideal datasets that may further contribute to the very issue we raise.

A limitation in this study is that the model used posed images. However, using scripts and dialogues in fictitious scenarios designed to elicit specific multimodal emotions within the context of discourse will enable realizations of more natural expressive interactions [4]. We are currently in the course of collecting such examples from non-actors and anticipate better gain in performance when our model is trained with this 'imperfect', yet natural, samples, which will be used in our future work. The future work will also implement domain adaptation techniques used in related works [2, 5, 7, 13, 14], in order to evaluate their effectiveness when applied to out-of-domain facial emotion training.

4. Conclusion

This study quantitatively verifies a limitation of the existing state-of-the-art facial emotion recognition models when applied to out-of-domain low quality images. We show its performance drops from 70.57% to 29.05%. This motivates our research to construct non-actor generated database, and the newly-trained model increases the accuracy by 17.81%.

Acknowledgments

This work was partly supported by the Korea government (2019-0-01367-BabyMind, 2017-0-01772-VTT, NRF-2017R1D1A1A02018479).

References

[1] Challenges in respresentation learning: Facial expression recognition challenge. https://www.kaggle.com/c/. *challenges-in-representation-learning-facial-expression-recognition-challenge*, 2013.

[2] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 233–242. JMLR. org, 2017.

[3] M. Biehl, D. Matsumoto, P. Ekman, V. Hearn, K. Heider, T. Kudoh, and V. Ton. Matsumoto and ekman's japanese and caucasian facial expressions of emotion (jacfee): Reliability data and cross-national differences. *Journal of Nonverbal behavior*, 21(1):3–21, 1997.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[5] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[6] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007.

[7] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[8] e. a. Jain, Neha. Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters 115*, pages 101–106, 2018.

[9] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1:39–58, 1997.

[10] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[11] D. Lundqvist, A. Flykt, and A. Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630, 1998.

[12] D. Matsumoto. Japanese and caucasian facial expressions of emotion (jacfee) and neutral faces (jacneuf). *Intercultual and Emotion Research Laboratory, Department of Psychology*, 1988.

[13] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

[14] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. *arXiv preprint arXiv:1812.01754*, 2018.

[15] C. Pramerdorfer and M. Kampel. Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*, 2016.

[16] M. A. Rouast, Philipp V. and R. Chiong. Deep learning for human affect recognition: insights and new developments. *IEEE Transactions on Affective Computing*, 2019.

[17] D. V. Sang and N. V. Dat. Facial expression recognition using deep convolutional neural networks. *9th International Conference on Knowledge and Systems Engineering (KSE)*, 2017.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. Casey, and C. Nelson. The nimstim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, 168(3):242–249, 2009.

[20] J. Van der Schalk, S. Hawk, and A. Fischer. Validation of the amsterdam dynamic facial expression set (adfes). *Poster for the International Society for Research on Emotions (ISRE), Leuven, Belgium*, 2009.