The 2nd Korean Emotion Recognition Challenge: Methods and Results

Songa Kim, Van Thong Huynh(✉), Dung Tran Thi, Aran Oh, Guee-Sang Lee,
Hyung-Jeong Yang, and Soo-Hyung Kim(✉)

Department of Artificial Intelligence Convergence, Chonnam National University,
Gwangju, South Korea
shkim@jnu.ac.kr

Abstract. The 2^{nd} Korean Emotion Recognition Challenge (KERC-2020) is a global challenge to promote the emotion recognition technologies by using audio-visual data analysis, especially for the emotion of Korean people. KERC2020 comprise of 1236 videos with each length from two to four seconds based on Korean movies are dramas. Around 68 participating teams compete to achieve state-of-the-art in recognizing stress, arousal, valence from Korean video in the wild. This paper provides a summary of dataset, methods and results in the challenge.

Keywords: Korean emotion recognition · Arousal · Valence · Stress · Multimodal · Affective computing

1 Introduction

Artificial intelligence is a computer program or a computer system that artificially implements human learning, reasoning, and understanding of natural languages. Simply, it is an artificial implementation of human intelligence on machines. The original purpose of various artificial intelligence researches was an experimental approach to psychology. However, the world is focusing on artificial intelligence, IoT, cloud computing, and big data in line with the fourth industrial revolution. The KERC aims to develop human emotion recognition technology in line with the original intentions of artificial intelligence research, especially by focusing on Korean emotion recognition to contribute to Korean emotion recognition research.

The KERC2020 decided stress on many emotional states as a topic. Stress means adaptation as a psychological and physical response that causes mental and physical stimulation. As many people in today's society are suffering from stress, we are trying to improve the quality of life and help people well-being by developing the technology to recognize people's stress. Furthermore, through KERC, we aim to increase the interest of Koreans in stress and emotion recognition technology.

H. Jeong and K. Sumi (Eds.): IW-FCV 2021, CCIS 1405, pp. 176–183, 2021.
https://doi.org/10.1007/978-3-030-81638-4_14

2 Dataset

KERC2020 dataset contain 1236 video clips with only one subject in each video. The collection process consists of cropping and remove low-quality videos. First, semi-automatic tool [7] was used to extract clips with a length of 2 to 4 s from 41 different Korean movies and dramas with various contexts. Then, we removed the low-quality clips such as obstructed face or the subject's back facing the camera. Each sample in the dataset is guaranteed that they focus on a clearly visible face, compose of facial expression of a subject in different activities. Table 1 describe some metadata information of our dataset.

Table 1. KERC2020 dataset metadata.

Attribute	Description
Scenario	In the wild
Source	Movie
Lengh of each sample	2–4 s
Frame resolution	$(720 \times 400) \sim (1920 \sim 1080)$ pixels
Number of samples	1236 (Training 636, Validation 300, Test 300)
Emotion categories	Stress, Valence, Arousal
Format	Video (mp4)

The label was annotated by 27 right-handed college students. They have no history of brain damage or psychiatric history, and currently do not take any medication. They were guided by the instructions that every reaction to a facial expression needs to be judged immediately and quickly as they feels it without need to worry or make a conscious effort to respond. The label annotation was performed in 2 days with 3 h each in the morning and afternoon. The students were divided into 2 groups of 14 and 13 people respectively. The first group annotated data on the first day's morning and second day's afternoon. The other group annotated on the remain time. The annotator were asked to rate each video clip on a 9-point scale from 1 to 9, which represent the low and high intensity of emotion. Each samples were annotated with 3 categories as in Table 2. Totally, ew have 33372 labels in each category for 1236 videos. The final score for each video (g) was obtained based on mean (μ) and standard deviation (σ) of scores from 27 annotators for that video as the following equation

$$g_c = \frac{\sum_{i=1}^{27} \alpha_{i,c} r_{i,c}}{\sum_{i=1}^{27} \alpha_{i,c}}, \tag{1}$$

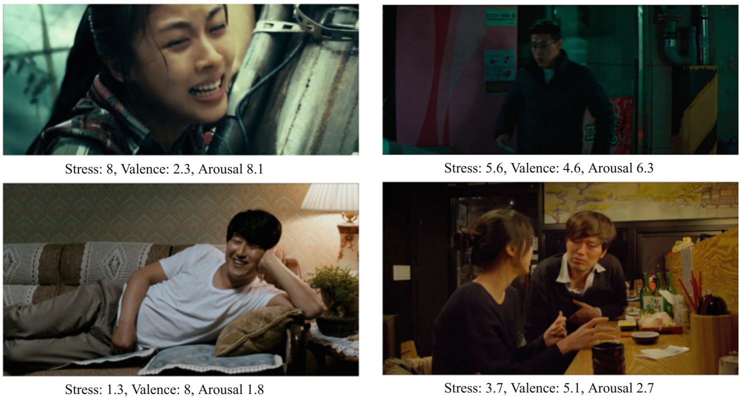
where $r_{i,c}$ is the score for emotion c which is rated by annotator i^{th}, and $\alpha_{i,c} \in \{0, 1\}$ indicate the using or elimination of the score from annotator i^{th} to reduce the dispersion of the data which is formulated as following

$$\alpha_{i,c} = \begin{cases} 1 & \text{if } \mu - 2\sigma \leq r_{i,c} \leq \mu + 2\sigma, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Table 2. The description of categories and its range in KERC2020 dataset.

Category	Description
Stress (1–9)	How stressed does a person feel? (Non-stressed ∼ stressed)
Valence (1–9)	How positive or negative an emotion is? (Negative ∼ Positive)
Arousal (1–9)	What is the agitation level of the person? (Inactive ∼ Active)

Figure 1 illustrated some examples of a frame in video clips of our dataset.


Stress: 8, Valence: 2.3, Arousal 8.1

Stress: 5.6, Valence: 4.6, Arousal 6.3

Stress: 1.3, Valence: 8, Arousal 1.8

Stress: 3.7, Valence: 5.1, Arousal 2.7

Fig. 1. Frame examples with video labels in KERC2020 dataset.

3 Baseline Approach

In this section, we describe our baseline method which is provided as a starting model for participants in KERC2020 challenge. Our approach consists of 3 stages: face detection, feature extraction, and score regression. In the first stage, we used Tiny Face Detector [6] to extract face region from any frames of each video, which produced 43328, 20314, and 20924 faces in training, validation, and test set, respectively. Each face is cropped and resized to 224 × 224 × 3 image in order to use as the input in second stage. We also resample each video to get 20 face image for each video before feeding to ResNet50 We deployed ResNet50 [5] architecture with pre-trained on VGGFace2, a large scale dataset for face recognition [2], as our feature extractor. We used the last average pooling layer of ResNet50 to obtain a feature vector of 20×2048 elements. In regression module, we deployed two LSTM layers followed by four fully connected layers. We built our baseline model on Keras and used Adam algorithm as optimizer with mean square error as objective function and learning rate of 0.001. A visualization of our approach can be seen in Fig. 2.

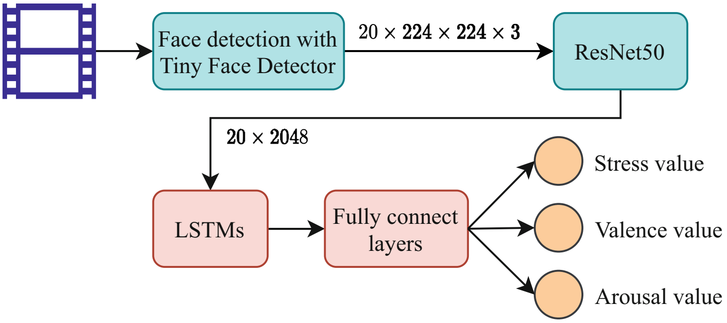


Fig. 2. The visualization of baseline architecture in KERC2020.

4 Challenge Methods and Results

The 2^{nd} Korean Emotion Recognition Challenge was hosted between August 20, 2020 and November 7, 2020 on Kaggle platform[1] which is used for downloading dataset and result submissions. The final ranking is based on private leader board which evaluated on test set and do not public to any participants until the end of challenge. Around 68 teams participated, with about 15 teams publicized result submissions. Submission were evaluated on the weighted average of three emotion categories, M, as following equations

$$\mathrm{M} = \frac{\mathrm{MSE}_{arousal} + \mathrm{MSE}_{valence} + 2 \times \mathrm{MSE}_{stress}}{4}, \tag{3}$$

where MSE indicate mean square errors. Table 3 shows the results of 2^{nd} KERC challenge. In this section, we review top 3 winner submission.

[1] https://www.kaggle.com/c/2020kerc/overview.

Table 3. Challenge results ranked by weighted average metric M.

Rank	Team name	Affiliation	Score (M)
1	Maybe Next Time	Chonnam Natl. Univ.	0.64838
2	pthmd	Chonnam Natl. Univ.	0.80898
3	scalable	Korea Univ.	0.81167
4	Han Soheon	Sungkyunkwan Univ.	0.93346
5	HouKM	Hallym Univ.	0.96578
6	iPsych	Korea Univ.(Empathy Research Institute)	1.47796
7	King Kong Intelligence	Korea Aerospace Univ.	1.58330
8	sswolf	–	1.58422
9	VI	Chosun Univ.	1.71953
10	sinu	–	1.73596
11	ISPL_emo	–	1.92627
–	**Baseline model**	–	**1.9283**
12	TT	–	1.92834
13	eep_learning	–	1.92834
14	emo	–	1.98704
15	Taeyoung Park	–	2.10328

4.1 Team Maybe Next Time

Their approach focus on the faces and leverage the emotion information from another facial expression datasets which included 3 stages: pre-processing, deep network regression, post-processing. In the first stage, the face region is detected and alignment with Multi-task Cascaded Convolution Networks (MTCNN) [13], then, a mask is used to remove forehead, hair, and anything outside the face. In the second stage, they used AffectNet dataset [10] and AFEW-VA dataset [8] to train the ImageNet pretrained model again. At this point, they fine-tune on 10 epochs with KERC2019 which is contained 7 discrete emotions KERC2020 dataset together in a multi-task scenario to leverage the relationship between continuous and discrete emotions. After that, in the last 5 epochs, they fine-tuned only on the KERC2020 dataset. Their predictions are in frame-level, then they averaged the results to obtain the final prediction for each video in the post-processing step. An illustration of their approach can be seen in Fig. 3.

4.2 Team Pthmd

They deployed an architecture which consists of two streams for audio and visual information. Each stream includes 2 stages: feature extraction, regression module. They leveraged pre-trained models on VGGFace2 [2] for visual information, and AudioSet [4] for audio datato extract the deep representation. Due to the varies in length of each sample, they performed average pooling to down-sampling the same time-dimension for all samples. At this point, PCA is used

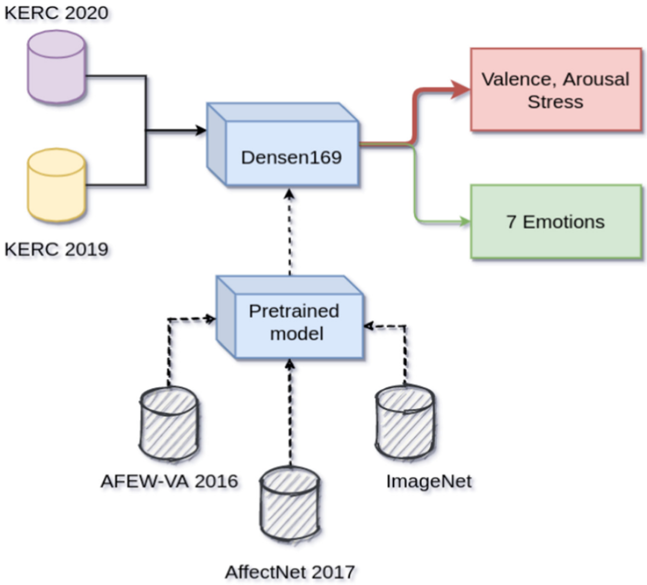


Fig. 3. Overview architecture of team Maybe Next Time.

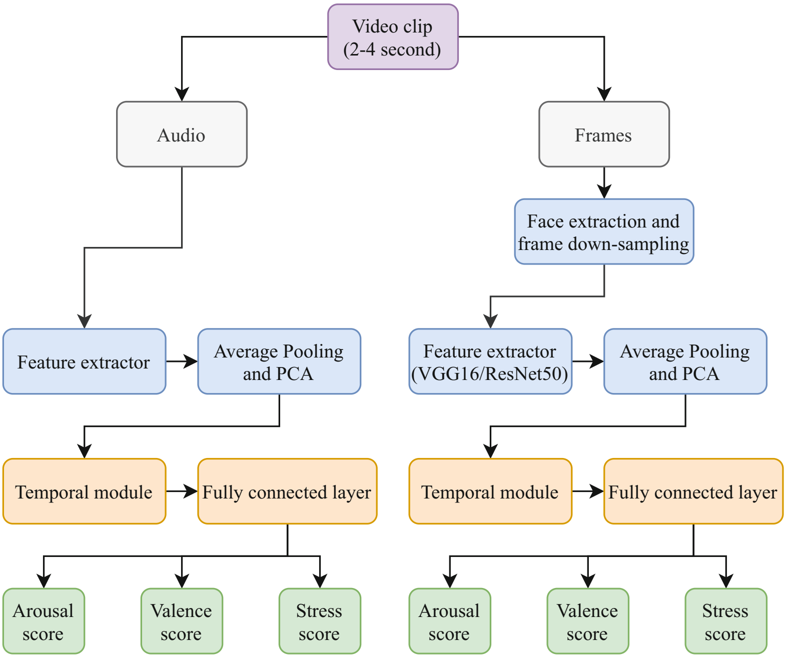


Fig. 4. An illustration of team pthmd's approach.

to select most emphasize features and used as the input to the next stage. In regression module, they deployed Temporal Convolutional Networks [1,11] to learn temporal relationship between frames instead of RNN based architecture to take advantage of parallelism, low memory training, and stable gradients. Then they use fully connected layers to obtain the final score for each emotion categories. Their best performance is achieved by the weighted average of results from different based feature extractor. Figure 4 show a visualization of their approach.

4.3 Team Scalable

They utilized Inception-ResNet-v2 [12] and Xception [3] as feature extractor. For visual information, they deployed both sequential model which involves LSTM layers, and frame-level model which average the results from each frame. They converted audio signals to logspectrogram, then fed them to the deep networks. They used Adam algorithm as optimizer and the learning rate is follow SGDR, a warm restart technique [9] to optimize their architecture. They achieved best performance with the ensemble of both audio and visual signals. Figure 4 show an illustration of their approach (Fig. 5).

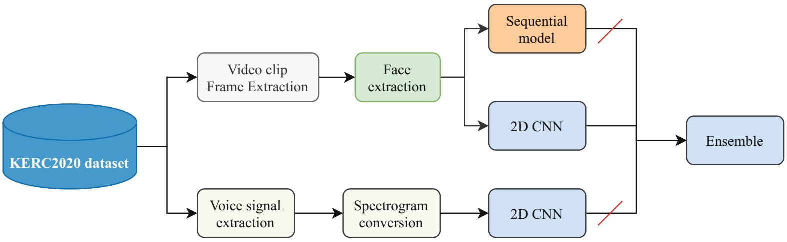


Fig. 5. An illustration of team scalable's approach.

5 Conclusion

Through the KERC2020, we promoted the development and interest of Korean emotion recognition technologies, and make a success. In particular, this competition focused on the topic of stress, especially for Korean people's stress. We provided participants with our dataset and baseline model to build and develop their own systems. As a result, various participants developed high performance methods. We will host the 3^{rd} KERC competition in this year of 2021 again to make a grater growth in the field of Korean emotion recognition.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A4A1019191) and by the Korea Sanhak Foundation and the University Industrial Technology Force (UNITEF) Support Group.

References

1. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, March 2018
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, May 2018. https://doi.org/10.1109/fg.2018.00020
3. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017. https://doi.org/10.1109/cvpr.2017.195
4. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, March 2017. https://doi.org/10.1109/icassp.2017.7952261
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2016. https://doi.org/10.1109/cvpr.2016.90
6. Hu, P., Ramanan, D.: Finding tiny faces. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017. https://doi.org/10.1109/cvpr.2017.166
7. Khanh, T.L.B., Kim, S.H., Lee, G., Yang, H.J., Baek, E.T.: Korean video dataset for emotion recognition in the wild. Multimed. Tools Appl. (2020). https://doi.org/10.1007/s11042-020-10106-1
8. Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.: AFEW-VA database for valence and arousal estimation in-the-wild. Image Vis. Comput. **65**, 23–36 (2017). https://doi.org/10.1016/j.imavis.2017.02.001
9. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts, August 2016
10. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2019). https://doi.org/10.1109/taffc.2017.2740923
11. van den Oord, A., et al.: WaveNet: a generative model for raw audio, September 2016
12. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-ResNet and the impact of residual connections on learning, February 2016
13. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016). https://doi.org/10.1109/lsp.2016.2603342