# A Multi-modal Approach for Emotion Recognition of TV Drama Characters Using Image and Text

Jung-Hoon Lee
*College of Computing*
*Sungkyunkwan University*
Suwon-si, South Korea
vhrehfdl@gmail.com

Hyun-Ju Kim
*College of Computing*
*Sungkyunkwan University*
Suwon-si, South Korea
julia981028@gmail.com

Yun-Gyung Cheong
*Department of AI*
*Sungkyunkwan University*
Suwon-si, South Korea
aimecca@skku.edu

***Abstract*—Research on facial emotion recognition has long been popular for various purposes. This paper investigates the recognition of the character emotions, to assist in understanding the story. The goal of this research is to classify the facial images of the characters in the Korean TV series 'Misaeng: The Incomplete'[1] into 7 emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. We built a multi-modal deep learning model which utilizes facial images as well as textual information describing the situations, to classify the facial images. Our experiments indicate that employing multi-modality enhances the performance of facial emotion recognition of story characters. We concludes with discussions and future work.**

***Index Terms*—Asian face, contextualized embedding, face emotion classification, multi-modal, CNN**

## I. Introduction

Facial emotion recognition has been studied for a long time and has been applied to sociable robotics, market research, facial emotion detecting smart cars, and job interviews. Recently, deep learning models such as Convolutional Neural Net have been applied to facial emotion recognition [1]. [2] surveys the state of the art (SOTA) of facial emotion recognition in widely known facial image datasets. The Extended CohnKanade (CK +) database classifies 593 image sequences into 7 emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise) and SOTA is 97.3% [3]. MMI classifies 1280 videos and over 250 images into 6 emotions excluding contempt (anger, disgust, fear, happiness, sadness, and surprise), and SOTA is 78.53% [4]. The FER2013 database was introduced in the International Conference on Machine Learning(ICML) 2013 Challenges,containing 28,709 training images, 3,589 validation images, and 3,589 test images. FER2013 classifies the images into 7 emotions (anger, disgust, fear, happiness, sadness, surprise and neutral) and SOTA is 75.2% [5]. MMI and FER2013 data show that external variables, such as a person's mustache or glasses, affect the performance, resulting in lower SOTA values.

The ultimate goal of our research is to develop an intelligent authoring tool for story generation, in which AI can assist a user to complete his or her story. When a story is fully written with the help of an AI storytelling technique, the authoring system pulls out appropriate video clips from the existing image pool, using the finished story. In order to find the appropriate footage for the story scenes, we need to automatically recognize the characters, objects, and activities in each of the scenes. We found out that facial emotions play an important role in finding and selecting appropriate footage to be used. For instance, when the story scene contains a couple talking to each other pleasantly, a footage of two people whose facial emotion contain fear cannot be used. Thus, the goal of the research described in this paper is to recognize the facial emotions of characters in a TV series. Our target video is the Korean TV series 'Misaeng: The Incomplete', which consists of 20 episodes. This dataset gives us two challenges. First, the characters are all Asian, while the models pre-trained with FER2013 and CK+ contain many Westerners' facial images. Second, the facial emotions of the main characters are expressed in a relatively subtle manner in the series. Due to these reasons, we conjectured that recognition of character emotion using the facial emotions alone can be limiting.

Therefore, we built face emotion data with 7 emotion classes (i.e., Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise) extracted from the TV drama series to train the recognition models. We also annotated each image with a textual description of the character's action.

In this paper, we built multi-modal deep learning models that recognize the expressed facial emotions of the characters in a scene. We compared the performance of models that trained with only images and those trained with multi-modals using both images and text.

The contributions of our research are as follows:

- We built image=text parallel training/test datasets that contain Asian characters' facial emotions annotated with a textual description of the action of the character.
- We built a multi-modal deep learning model using image and text information.

This paper sketches the related work in section 2. Section 3 describes the structure of the data and how it is created. Section 4 describes our proposed method that uses image and text for emotion classification. In section 5, we explain our experiments and results. Finally, we conclude with discussions and future work in section 6.

[1] http://program.tving.com/tvn/misaeng/

2375-9356/20/$31.00 ©2020 IEEE
DOI 10.1109/BigComp48618.2020.00082

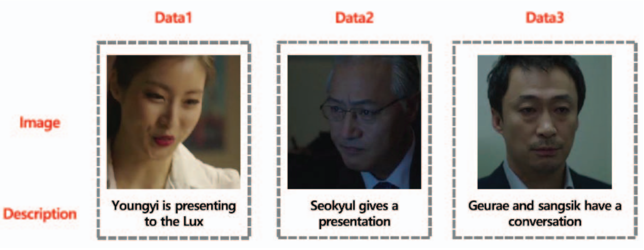

Data1
Data2
Data3
Image
Description
Youngyi is presenting to the Lux
Seokyul gives a presentation
Geurae and sangsik have a conversation

Fig. 1. Samples facial emotions from Misaeng.

## II. Related Work

Emotion recognition has been investigated by various disciplines such as psychology, sociology, cognitive science, and computer science. Based on the most well-known works about emotion [6], [7], the basic classes of emotion are happiness, sadness, anger, surprise, fear, contempt, and disgust. In order to recognize these emotion categories, researchers have employed various input features: visual data such as images and videos [1], audio [8] or texts [9]. In Computer Vision, previous research works have made efforts to develop features like SIFT (Scale Invariant Feature Transform). As deep learning techniques have advanced, Convolutional Neural Networks (CNNs) are generally used to extract features from images. Recently, very deep learning architectures like VGGNet, ResNet have shown the best results in image classification problems. End-to-end models using CNN and attention-based [10] models are also promising.

For processing textual data, there are various works analyzing words to sentences Natural Language Processing (NLP) research. To encode each word's semantic meaning, distributive embedding deep learning models (e.g., Word2Vec [11], Glove [12], BERT [13], ELMo [14]) are used to classify emotions of a given text. Word2vec, Fasttext [15], and Glove produce fixed vector values for each word. On the other hand, the contextualized embedding such as ELMo and BERT method can produce the vector value of a word based on its position and surrounding words.

Instead of extracting features from only one type of data sources, combination of different input can be used, such as combination of audio and text [16], [17], or audio and video [18].

The well-known benchmark datasets for emotion recognition are FER, JAFFE, and FERG. When we surveyed widely used datasets, we found that most of the image data used for detecting facial emotions were based on Westerns' facial images, with the exception of JAFFE. When the machine learning models for emotion recognition trained with Westerners' facial images, the application of the model on Asians' facial images tend to show lower performance [19].

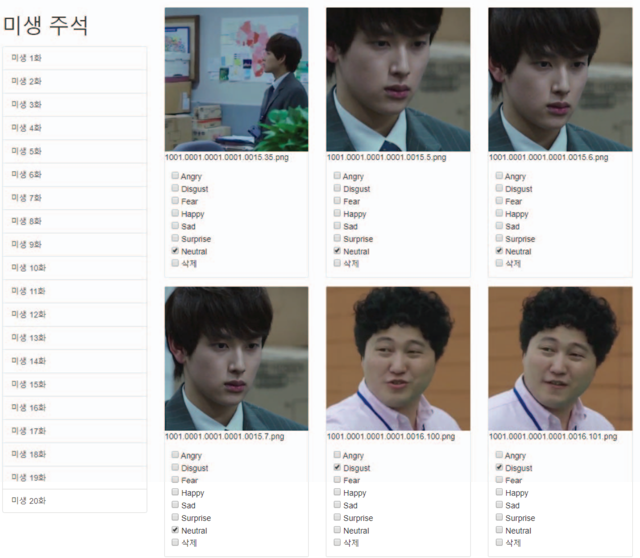
## III. Data

Our dataset is built based on the TV series 'Misaeng: The Incomplete', where all the characters are Asian. The model pre-trained with FER2013[2] data was applied to a drama dataset, and the accuracy score was less than 40%. In fact, face recognition has shown that different races of facial image datasets have some effect on recognition performance [19].

We construct a dataset by pairing an image and text as shown in Figure 1. The image in Figure 1 contains the face of the characters in the video, and the text is the description that describes the situation in which the image appears. The description consists of the characters and actions in the scene as some examples listed below.

- Geurae jumps.
- Geurae makes a phone call.
- Geurae and Sangsik have a conversation.
- Geurae and Seokyul give a presentation.
- Geurae, Dongsik, Sangsik and Mechelle are in a meeting.

To built the data, we used FFMPEG[3] to cut 10 one hour-long episodes into 1-second frames to build an image dataset. And using OpenCV[4], the facial image in each image was extracted. The cropped facial images were uploaded to a web page and manually classified into 7 classes of emotion. Figure 2 is the screen of the website.



1001.0001.0001.0001.0015.35.png
□ Angry
□ Disgust
□ Fear
□ Happy
□ Sad
□ Surprise
☑ Neutral
□ 삭제

1001.0001.0001.0001.0015.5.png
□ Angry
□ Disgust
□ Fear
□ Happy
□ Sad
□ Surprise
☑ Neutral
□ 삭제

1001.0001.0001.0001.0015.6.png
□ Angry
□ Disgust
□ Fear
□ Happy
□ Sad
□ Surprise
☑ Neutral
□ 삭제

1001.0001.0001.0001.0015.7.png
□ Angry
□ Disgust
□ Fear
□ Happy
□ Sad
□ Surprise
☑ Neutral
□ 삭제

1001.0001.0001.0001.0016.100.png
□ Angry
☑ Disgust
□ Fear
□ Happy
□ Sad
□ Surprise
□ Neutral
□ 삭제

1001.0001.0001.0001.0016.101.png
□ Angry
☑ Disgust
□ Fear
□ Happy
□ Sad
□ Surprise
□ Neutral
□ 삭제

Fig. 2. Website that categorizes images into 7 emotions (Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise)

## IV. Our Approach

This section describe our approach. For training and evaluation, we use TV series 'Misaeng: The Incomplete' (released in 2014), where the story depicts various situations in the work place. One problem particular to our data is that the facial emotions conveyed by the characters are very subtle. The main character of the series is a rookie intern named Geu-rae Jang (played by Si-wan Im), who is a very shy introvert. In order

[2] https://www.kaggle.com/deadskull7/fer2013
[3] https://www.ffmpeg.org/
[4] https://opencv.org/

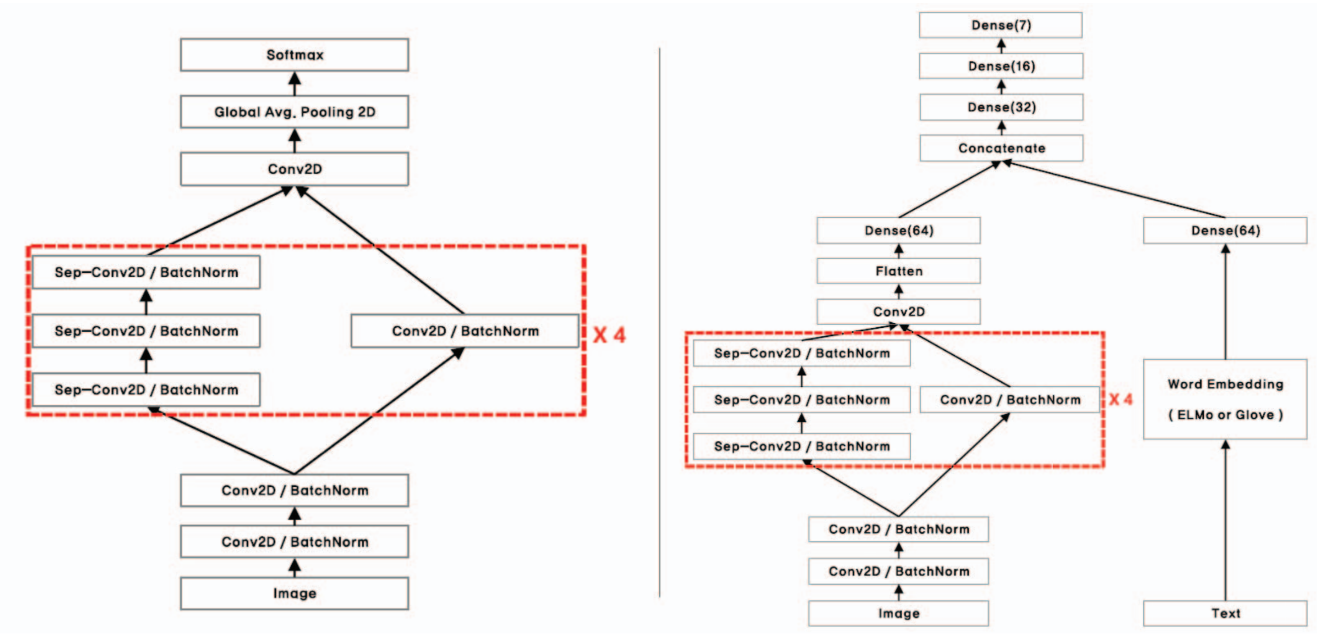


Fig. 3. Our deep learning model architecture. On the left, a CNN-based model that uses only facial images of characters as input data. There are 7 outputs by activating in Global Average pooling layer without dense layer. The largest of the output values is selected as the emotion class of the image. On the right (Multi-modal): a model that adds ELMo embedding to Base-model. In multi-modal, input data is inputted with image and description of the character's action. Finally, we concatenate the image and text vector values and apply softmax to produce 7 output values.

to become a full-time employee, the protagonist rarely shows strong emotions such as anger or joy. This complicates the classification of emotions. To solve this problem, we built a multi-modal deep learning model that uses a character's facial images and text describing the situation as input values.

### A. *Multi-modal vs Single-modal*

We hypothesized that using textual description would improve the classification performance when facial emotion is too subtle to convey clear emotions of a character. To test the hypothesis, we compared the performance of a baseline model trained using only images with the performance of a multi-modal model trained using both images and text. As the input data types, we use facial images of the characters and text describing the situation in the given scene.

The baseline model uses the Convolutional Neural Networks based model, called 'Sequential fully-CNN', proposed in [1]. The overall architecture of the model is illustrated on the left side in Figure 3. The Sequential fully-CNN model, consisting of 9 convolutional layers, uses a softmax activation function for each reduced feature map. The number of feature maps in the last convolutional layer is composed of 7 classes. As the optimizer, Adam and RELU are used as the activate function. For regularization, Batch Normalization method and Global Average Pooling method are used.

The overall architecture of the multi-modal is shown in the right part of Figure 3. The image processing part of the multi-modal is based on the sequential fully-CNN, which is the baseline model. When the image is flattened through the sequential fully-CNN, the image is output in 2527 dimensions and compressed into 64 dimensions through the dense layer.

The text is inputted into the input layer and converted into a vector via the embedding layer. It is compressed into 64 dimensions through the dense layer. The multi-modal model concatenates the compressed image and text into 64 dimensions to create 128 dimensions, which are then compressed into 7 dimensions after going through multiple dense layers. In the output layer, the largest of the 7 output values are selected as the emotion for the facial image. The multi-modal model also uses the Adam optimizer and the RELU function.

### B. *Word Embedding*

In order to apply the text to the deep learning model such as CNN and LSTM, we need to convert it to a vector. In this study, we also test the performance of different word embedding methods: Glove and ELMo. We chose these two embeddings models because Glove employs the co-occurrences of the whole corpus and the pre-trained model has learned general meanings from a variety of large corpora. In this paper, we use the pre-trained 300-dimensional Glove embedding vector[5]. The model converts words into 300 dimensions vector using Glove and compresses them into 64 dimensions vector by applying a convolution mask with filter sizes of 3, 4 and 5 [20].

ELMo is chosen because it takes less time to train the model due to a small number of parameters compared with other models such as BERT. We use the pre-trained weights provided

[5] https://nlp.stanford.edu/projects/glove/

TABLE I
A SIMPLE ANALYSIS OF DATA

| Emotion | Train | Test | Total |
|---|---|---|---|
| Angry | 1,442 | 386 | 1,828 |
| Disgust | 472 | 114 | 586 |
| Fear | 135 | 41 | 176 |
| Happy | 828 | 193 | 1,021 |
| Neutral | 1,457 | 343 | 1,800 |
| Sad | 891 | 218 | 1,109 |
| Surprise | 585 | 158 | 743 |

TABLE II
EXPERIMENT RESULT.(F1 SCORE)

| Emotion | Base-model | Multi-modal(Glove) | Multi-modal(ELMo) |
|---|---|---|---|
| Angry | 0.54 | 0.63 | **0.64** |
| Disgust | 0.07 | **0.42** | 0.39 |
| Fear | 0.0 | 0.19 | **0.21** |
| Happy | 0.52 | **0.63** | 0.6 |
| Neutral | 0.56 | **0.69** | **0.69** |
| Sad | 0.47 | 0.52 | **0.53** |
| Surprise | 0.07 | 0.43 | **0.46** |

by TensorFlow Hub[6] and fine-tuned our data to improve performance [21]. Hence, we built two different types of multi-modal models using Glove and ELMo as the embedding layer.

## V. EXPERIMENTS AND RESULTS

### A. Experiments

A total of 7,263 images were created and converted in gray scale at 300 x 300 pixel size. The counts of images for each emotion category is shown in Table I. We trained each model with 100 epochs. The data were randomly divided into training and test sets at an 8: 2 ratio.

### B. Result and Discussion

We compare the recognition performances of the three models: A baseline model, the Glove multi-modal model, and the ELMo multi-modal model.

As in Table II, both multi-modal models show higher recognition performance than the baseline model in terms of the F1 score. In particular, the base-model performed poorly for the disgust, fear, and surprise classes. Significant differences in performance were observed in the surprise class (difference of 0.39) and in the disgust category (difference of 0.35). The multi-modal(ELMo) resulted in a low performance, with an F1 score of 0.21 in the fear category. However, with the baseline model, no images were classified as fear, due to class imbalance in the training set. The highest F1 score of 0.69 was obtained for the neutral category when using the ELMo multi-modal model.

Overall, the multi-modal (ELMo) model showed high performances in four categories : angry, fear, sad and surprise. The multi-modal (Glove) model showed high performances in three categories: disgust, happy, and neutral. Both achieved the highest performance in the neutral class. The difference between these two models are marginal.

These results indicate that textual description can enhance the classification performance significantly, particularly in the context where emotions are not clearly identified through facial images.

## VI. CONCLUSION

In this paper, we describe our multi-modal deep learning method to classify facial images into 7 emotional categories (angry, disgust, fear, happy, neutral, sad, surprise). For this end, we built a dataset that contains facial images. The data has three problems. First, it contains Asian facial expressions, which have been little discussed. Second, the emotion categories are imbalanced. Finally, the emotions were expressed subtly on the images. To address these problems, we built two multi-modal models for classifying emotions using images and text. Our experiment results suggest that using text description of the characters' actions enhances the recognition performance significantly.

In the future, we plan to augment data for rarely represented classes (e.g., disgust and fear) using generative deep learning models such as Generic Adversarial Networks (GAN). We also plan to use the state of the art in word embeddings such as BERT and XLNet.

## ACKNOWLEDGMENT

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2019R1A2C1006316) and Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-00421, AI Graduate School Support Program). This research was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01642) supervised by the IITP(Institute for Information & communications Technology Promotion)

## REFERENCES

[1] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018.

[3] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168–2177.

[4] X. Liu, B. Vijaya Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–29.

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

[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[7] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[8] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," *arXiv preprint arXiv:1912.00846*, 2019.

[6] https://tfhub.dev/google/elmo/2

[9] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24–35, 2018.

[10] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *arXiv preprint arXiv:1902.01019*, 2019.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[16] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *arXiv preprint arXiv:1911.00432*, 2019.

[17] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[18] Z. Lian, Y. Li, J. Tao, and J. Huang, "Investigation of multimodal features, classifiers and fusion methods for emotion recognition," *arXiv preprint arXiv:1809.06225*, 2018.

[19] Z. Xiong, Z. Wang, C. Du, R. Zhu, J. Xiao, and T. Lu, "An asian face dataset and how race influences face recognition," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 372–383.

[20] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[21] C. S. Perone, R. Silveira, and T. S. Paula, "Evaluation of sentence embeddings in downstream and linguistic probing tasks," *arXiv preprint arXiv:1806.06259*, 2018.