Korean video dataset for emotion recognition in the wild

Trinh Le Ba Khanh¹ · Soo-Hyung Kim¹ · Gueesang Lee¹ · Hyung-Jeong Yang¹ · Eu-Tteum Baek¹

Received: 14 October 2019 / Revised: 2 September 2020 / Accepted: 19 October 2020 /
Published online: 10 November 2020
© The Author(s) 2020

Abstract

Emotion recognition is one of the hottest fields in affective computing research. Recognizing emotions is an important task for facilitating communication between machines and humans. However, it is a very challenging task based on a lack of ethnically diverse databases. In particular, emotional expressions tend to be very dissimilar between Western and Eastern people. Therefore, diverse emotion databases are required for studying emotional expression. However, majority of the well-known emotion databases focus on Western people, which exhibit different characteristics compared to Eastern people. In this study, we constructed a novel emotion dataset containing more than 1200 video clips collected from Korean movies, called Korean Video Dataset for Emotion Recognition in the Wild (KVDERW). Which are similar to real-world conditions, with the goal of studying the emotions of Eastern people, particularly Korean people. Additionally, we developed a semi-automatic video emotion labelling tool that could be used to generate video clips and annotate the emotions in clips.

Keywords Emotion recognition · Facial expression · Emotion dataset

1 Introduction

Emotion recognition systems scientifically measure and analyze complex feelings, such as comfort, discomfort, convenience, and inconvenience, in humans to guide product or environment design to improve the quality of human life. Practically speaking, emotion recognition technology can facilitate emotion-based services for users by detecting emotions in entertainment, education, medicine, etc. This technology enables the analysis of immediate user reactions at the time of service, thereby improving the quality of services [11].

Emotion recognition has attracted significant attention from the computer vision and affective computing communities over the past decade because it is an important front-end task in many applications. The majority of existing techniques focus on classifying seven

✉ Hyung-Jeong Yang
hjyang@jnu.ac.kr

¹ Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea

basic expressions: *anger*, *disgust*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise* [8, 18]. A few methods follow a dimensional approach in which emotional expressions are treated as regression data in the arousal-valence space [16, 20].

Recently, deep learning approaches have been proposed to classify emotions [12] [14] [1]. Based on the outstanding performance of deep learning approaches, emotion recognition models have achieved remarkable accuracy. However, emotion recognition is still a very difficult task for deep learning architectures. In particular, one of the major limitations of emotion recognition is a lack of appropriate emotion databases.

The majority of existing emotion databases focus on Western people, whose tone and energy of speech, facial expression, lip-reading, and gaze are slightly different during natural communication in comparison with Eastern people. Benitez-Garcia et al. [2] presented a methodical analysis of Western and Eastern prototypical expressions with a focus on four facial regions: forehead, eyes-eyebrows, mouth and nose. They determined that the major differences between Western and Eastern facial expressions occur in the regions of the mouth and eyes-eyebrows, specifically for expressions of *disgust* and *fear*. Similarly, the authors [10] report marked differences between Eastern and Western observers in the decoding of universal facial expressions. Eastern observers exhibited an inadequate to reliably distinguish universal facial expressions of *fear* and *disgust*. Western observers distributed their fixations evenly across the face, whereas Eastern observers persistently fixate the eye region. Additionally, emotion is not only biological, but is also influenced by the environment, which differs significantly between Western and Eastern culture [13]. Therefore, datasets with a focus on Eastern people are required for studying emotion recognition in Eastern cultures.

In this paper, we propose a novel emotion video database containing over 1200 video clips of Eastern people (mostly Korean people). Our database contains video clips collected from 41 Korean movies. The clips were annotated by six Korean evaluators. The database has a main face designated in every frame of each clip to provide more clear information about the object for recognition. Each clip has clear emotion conditions to facilitate easy for training and the development of emotion recognition models. This dataset provides information about the changes in face and background. This is appropriate for models about emotion recognition using facial expression, head pose and background. In addition, it provides visual and audio for studies using multimodal. We also propose a semi-automatic authoring tool to support the rapid creation and labelling of emotion video clips. This authoring tool can automatically cut an original video into a list of clips for a particular face. It also provides default emotion prediction for clips.

The remainder of this paper is organized as follows. Section 2 briefly describes well-known datasets for emotion recognition. Section 3 introduces the proposed video labelling tool and presents the detailed process used for generating and annotating the proposed emotion database. Finally, we present a baseline model for evaluating the proposed dataset in Section 4 and discuss the conclusions of this study in Section 5.

2 Related work

Emotion recognition methods can be categorized based on two types of environments, namely “lab controlled” and “in the wild”. Most human facial expression databases and methods largely depend on lab-controlled environments. This implies that most existing databases were created by controlling backgrounds, illumination, and head movement, indicating that they are not representative of real-world scenarios. We compare KVDREW dataset with other emotion recognition datasets such as MMI [15], IEMOCAP [3], AFEW [5], Belfast

Table 1 Comparison of the KVDERW with existing emotion recognition datasets such as AFEW [5], Belfast [7], MMI [15], IEMOCAP [3]

Dataset	Process	Amount of data	Setting	Multiple subjects	Region
KVDERW	Assisted	1200 clips	Movie	No	Eastern
MMI [15]	Manual	1280 clips	Lab	No	Western
IEMOCAP [3]	Manual	7500 clips	Lab	No	Western
AFEW [5]	Assisted	1809 clips	Movie	Yes	Western
Belfast [7]	Manual	239 clips	TV & Lab	No	Western

[7] as shown in Table 1. The MMI [15] database is a facial expression database containing both images and videos of 75 subjects in a lab-controlled environment. This database does not accurately capture the conditions of real-world scenarios. The IEMOCAP [3] dataset is one of the most commonly used databases for researching emotion recognition systems. Although this dataset is used by many researchers, it is heavily limited by data collection in a lab-controlled environment.

Based on the limitations of natural human expression in lab-controlled environments, several databases and methods representing close-to-real-world environments have been proposed recently. These databases cover unconstrained facial expressions, different age ranges, different face resolutions, varying focus levels, and real-world illumination conditions. The AFEW [5] dataset is a facial expression database consisting of close-to-real-world conditions from 54 movies with data collected using a recommender system based on subtitles. This system extracts subtitles containing expression-related keywords. The length of each video clip is equal to the time of appearance of the corresponding subtitle. The video clips were annotated with seven basic expressions, namely *anger*, *disgust*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*, by human labelers. The database contains 330 subjects with a large age range of 1–77 years and addresses the issue of temporal facial expressions in the wild. The SFEW [4] dataset was created by selecting specific frames from the AFEW [5] dataset. The SFEW [4] dataset contains 700 images labeled with seven basic emotions that were annotated by two labelers. The AFEW [5] dataset consists of multiple labeled subjects in the same frame, therefore, it don't provide rich information of target object for training such as the changes in face expression and head pose. The Belfast [7] database consists of 239 clips collected from TV programmes and the interview recordings labeled with particular expressions. The number of TV clips in this database is sparse. Compared to the manual method which was used to construct and annotate, our labelling tool is faster and more effective.

Recently, deep learning technology has advanced significantly and achieved remarkable results. Therefore, the organized challenges related to automatic emotion recognition have grown in popularity. Two of the most difficult recent challenges are the Emotions in the Wild (EmotiW) challenge [6] and Aff-Challenge [19]. Both challenges use in-the-wild datasets. EmotiW uses clips from movies annotated based on six universal emotions and a *neutral* emotion. AFF-Challenge uses YouTube videos annotated based on dimensional arousal and valence scales. These datasets include extremely difficult conditions for recognizing the emotions of targets.

3 Video labelling tool and database creation procedure

3.1 Video labelling tool

The proposed semi-automatic labelling tool was designed to support the creation of datasets and annotations. The authoring tool was developed using the C # Windows Presentation

Fig. 1 User interface for the proposed semi-automatic authoring tool for creating and annotating emotion video clips

Foundation framework. It provides an application interface that can easily create video clips and annotate emotions. The recorded annotations can be exported in the comma-separated variables (.csv) format for data analysis via majority voting or multiple labels.

The proposed labelling tool includes two main tasks: automatic cutting of an original video into a list of video clips based on faces and annotation of emotion labels by multiple annotators. Figure 1 presents the user interface of the proposed labelling tool with imports indexed by number. Box one is the interface for cutting an original video. Box two contains control buttons for video playback. Boxes three and four contain interfaces for annotating emotion labels based on categories and continuously, respectively. Box five presents information about the current video. In the following section, we describe our data generation procedure.

3.2 Data generation

The process of creating and annotating emotion video clips has three main steps. First, we create a list of video clips by automatically cutting an original video based on a face identification process. This process is performed using the interface in box one in Fig. 1. Next, we eliminate low-quality video clips from the list of clips, as shown in Fig. 3. Finally, evaluators annotate emotions in video clips. To obtain a final emotion label, one can use either majority voting or multiple labels for the process in box three in Fig. 1. In the real world, an emotion can be recognized differently based on the target individual. In this study, only one emotion label was assigned for each video clip. Multiple labels will be examined in future research. Figure 2 presents an overview of the procedure of creating and annotating emotion video clips.

Fig. 2 Flow chart illustrating the process of creating and annotating emotion video clips

3.2.1 Cutting the original video

Cutting an original video into clips manually requires significant time and effort. To handle this problem, we designed an algorithm that automatically extracts video clips from an original video. Video clips must contain at least one face in every frame. A video clip is defined as a set of consecutive frames containing the same face.

For detecting faces in videos, we integrated several face detection algorithms, including OpenCV, Dlib, Mtcnn [21], and Tinyface [9], as shown in the face detection section of box one in Fig. 1. The proposed interface allows users to select the desired algorithms for face detection. To determine if the faces in consecutive frames represent the same person, we calculate the Euclidean distances between face features extracted by a pre-trained VGG16 model [17] based on the VGGFace dataset. In the proposed system, cutting one video into clips takes approximately 10 hours when using Mtcnn face detection with a clip length of 2 seconds. The video cutting algorithm is outlined in Algorithm 1.

To test the cutting algorithm, approximately fifty thousand video clips from 41 movies were generated. Most clips were eliminated in the deletion step to generate a final dataset with good conditions for recognizing emotions.

Algorithm 1 Cutting an original video.

Input : Original video

Output: List of video clips

Setting: Length of video clip: l (number of frames in a video clip)

Step 1:

1 **if** the current frame index (c) has the largest greatest proportion of for one face (f) **then**

2 update the position of the beginning frame index(s) for cutting: $s := c$;
3 move to the next frame: $c := c + 1$;

4 **else**

5 move to the next frame: $c := c + 1$;
6 repeat Step 1 ;

Step 2:

7 **if** the largest face proportion (f) still exists in current frame index (c) **then**

8 update the current face number (n) of f : $n := n + 1$;
9 move to the next frame: $c := c + 1$;

10 **else**

11 delete the position of the beginning frame index (s) for cutting ;
12 move to the next frame: $c := c + 1$;
13 repeat Step 1 ;

Step 3:

14 **while** $n < l$ **do**

15 repeat Step2;

16 **if** $n == l$ **then**

17 update the position of ending frame index (e) for cutting: $e := c$;
18 cut a sub video clip from frame s to frame e of the original video ;
19 move to next frame: $c := c + 1$;
20 repeat Step 1 ;

Fig. 3 Examples of low-quality video clips

3.2.2 Delete low-quality video clips

The quality of the video clips created automatically by the tool will vary significantly. One of the most common bad conditions is when the same content is repeated in consecutive video clips. In other words, there are many similar video clips. We eliminate as many similar video clips as possible to enhance the quality of dataset. Furthermore, our dataset consists of video clips focusing on facial expression, therefore, the bad condition is the appearance of factors affecting the quality of video clips for the emotion recognition, such as part of a face disappearing, illumination variation on a face, incorrectly identified faces, faces being covered by helmets or masks, faces appearing behind glass or a fence, and faces being covered by blood or color. Therefore, the face detection is hard to recognize the seven emotions. We remove such video clips to obtain a final set of video clips that is suitable for the seven emotion recognition. Figure 3 presents some examples of low-quality video clips for the emotion recognition generated by the proposed tool.

The final dataset used in this study contained 1316 video clips with good conditions for recognition. These clips were manually annotated to get the final emotion labels. The process of annotating was described in the following session.

3.2.3 Categorical emotion annotation

Six native Korean evaluators (two female and four male university students) were asked to annotate the emotional content in the database based on seven emotional categories. The database was divided into 41 folders corresponding to 41 movies. Each folder was annotated in sequence. Six evaluators assigned labels separately for each folder. Each folder was annotated a single session. They were restricted from seeing the annotations of other evaluators to avoid any bias. No discussion between evaluators was allowed. The evaluators were

Table 2 Parameters for video clip annotation

Attribute	Description
Number of annotators	6
Annotation time per clip	20–40 (s)
Number of labels each annotator could assign	1
Required votes to assign a final label	Agreement between at least four people

Fig. 4 Distribution of emotional categories, where the total number of video clips is 1316

instructed to take a suitable rest between annotating folders. The database¹ was designed to target seven categorical emotions: *anger*, *disgust*, *fear*, *happiness*, *neutral*, *sadness*, and *surprise*. The process of annotation required approximately 12 hours and was conducted in three sessions over three days, with four hours of annotating per day.

The evaluators were asked to select only a single emotion label for each video clip. For the sake of simplicity, majority voting was used for emotion assignment with a label requiring at least four votes to be selected as the final label. Table 2 summarizes the parameters for clip annotation. Figure 4 and Table 3 present the distribution of emotion labels assigned by at least four people, where the *other* label in Fig 4 indicates labels assigned by less than four people. One can see that the emotion database has a balanced distribution of target emotions overall, but the *disgust* emotion has relatively few samples. The total number of video clips in Table 3 is 1246, which excludes the other category in Fig 4.

Table 4 summarizes the metadata of the proposed database. Figure 5 presents some example emotions from the proposed database, which are very similar to real world scenarios.

4 Baseline experiments

In this section, we present baseline experiments that were performed to determine the quality of the proposed dataset. The dataset was divided into training (50%), validation (20%), and testing sets (30%) consisting of 639, 271, and 336 clips, respectively.

To predict the emotion labels of video clips, we first extract the main face from each frame in a clip using the Tinyface detector. Figure 6 depicts an example of the main face in

¹<https://www.kaggle.com/c/kerc/data>

Table 3 Emotion distribution for KVDERW

Emotion	Number of Clips	Percentage of labels
Anger	354	28.41%
Disgust	34	2.73%
Fear	123	9.87%
Happiness	207	16.61%
Neutral	73	5.86%
Sadness	273	21.91%
Surprise	182	14.61%
Total	1246	100%

each frame of a video clip. The list of main faces is then down- or up-sampled to generate a final list containing 20 faces for each video clip. Features are then extracted from each face in the list using a pre-trained VGG16 model [17] based on the VGGFace dataset. The face features from each face are concatenated to create a single feature vector. Finally, a multi-layer perceptron (MLP) classifier is trained to classify video emotions. During training, the weights of the pre-trained models were frozen, implying that only the weights of the MLP were adjusted. We used the Adam optimizer for training with a learning rate of $1e-6$ and batch size of 8. Figure 7 illustrates the pipeline for our baseline model.

The baseline model obtained accuracies of 44.28% and 39.58% on the validation and testing sets, respectively. Figures 8 and 9 present the confusion matrices for the baseline model on the validation and testing sets, respectively. The model performed well on the *anger*, *happiness*, and *surprise* emotions. In contrast, the *disgust*, *fear*, *neutral* categories are difficult to recognize. It may be caused by the imbalanced distribution of the data for different categories. From the data distribution shown in Table 3, the label number ratios for these three categories are less than 10% for each category in the entire dataset (*disgust* (2.73%), *fear* (9.87%) and *neutral* (5.86%)). While the *anger*, *happiness*, *sadness* and *surprise* emotions categories compose most of the dataset.

Additionally, we created an English emotion dataset by collecting video clips from 23 movies. The pipeline for creating this dataset is the same as that for creating KVDERW. This dataset consists of 925 clips. The emotion label distribution for this dataset is presented in Table 5. The English video dataset was divided into training (50%), validation (20%), and testing (30%) sets consisting of 489, 192, and 244 clips, respectively. We conducted

Table 4 Dataset metadata

Attribute	Description
Scenario	In the Wild
Length of video	2–4(s)
Size of frame	(720x400)–(1920x1080) (px)
Number of videos	1246
Expression classes	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise
Annotation Level	Video (mp4)

Fig. 5 Example emotions in database. From left to right, top to bottom: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise

Fig. 6 Example of main face in each frame of a video clip

Fig. 7 Pipeline for the baseline model for emotion recognition

Fig. 8 Confusion matrix for the Korean validation set

Fig. 9 Confusion matrix for the testing set of KVDERW

an evaluation on this dataset using a baseline model trained in the same manner discussed above. The confusion matrices for the validation and testing sets are presented in Figs 10 and 11, respectively, with accuracy values of 43.23% and 37.71%, respectively. One can see that the *fear* and *sadness* emotions are predicted accurately, whereas *disgust* and *neutral* are often misclassified.

We also tested a baseline model trained using KVDERW on the English testing set and vice versa. The experimental results are listed in Table 6. One can see that there are differences between the confusion matrices for the testing set of KVDERW presented in Figs. 9 and 12, and between the confusion matrices for the English testing set presented in Figs. 11 and 13. For example, the *fear* emotion is more prominent in Fig. 9 than in Fig. 12. Similarly, the *fear* emotion is accurately predicted in Fig. 11 and poorly predicted in Fig. 13. These large differences demonstrate that there are differences between the expressions of emotions of Eastern and Western people. Similar to the *disgust*, *fear*, *neutral* categories in

Table 5 Emotion distribution for the English dataset

Emotion	Number of Clips	Percentage of labels
Anger	195	21.08%
Disgust	14	1.51%
Fear	213	23.03%
Happiness	104	11.24%
Neutral	33	3.57%
Sadness	243	26.27%
Surprise	123	13.30%
Total	925	100%

Fig. 10 Confusion matrix for the English validation set

Fig. 11 Confusion matrix for the English testing set

Table 6 Experiments on different testing sets

Training set	Validation set	Testing set	Accuracy on testing set
Korean	Korean	Korean	39.58%
English	English	Korean	26.79%
Korean	Korean	English	34.43%
English	English	English	37.71%

Fig. 12 Confusion matrix for the testing set of KVDERW with pre-training on the English set

KVDERW, the results from three categories the *disgust*, *neutral*, *surprise* categories are not good. It may be caused by the imbalanced distribution of the data for different categories (shown in Table 5) and the limit data size.

Fig. 13 Confusion matrix for the English testing set with pre-training on KVDERW

5 Conclusion

In this study, we proposed a novel Korean emotion dataset containing emotion video clips under close-to-real-world conditions. The clips were extracted from movies for emotion recognition and the database was assessed using a baseline system. The main purpose of this dataset is to help researchers more easily develop and improve emotion recognition models for various races. To reduce costs and save time when creating and labelling videos, we also developed a semi-automatic tool for supporting the creation and annotation of emotion video clips. This type of tool can help users identify difficult cases when assigning emotion video labels by providing clearer information regarding a video clip as a whole. In the future, we will improve this tool and consider an additional dataset based on the arousal-valence space.

Acknowledgements This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (NRF-2020R1A4A1019191). The authors also thank Ms Tran Thi Dung for participation in revising of the manuscript. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2016-0-00314) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Avola D., Cinque L., Fagioli A., Foresti G. L., Massaroni C. (2020) Deep Temporal Analysis for Non-Acted Body Affect Recognition. In: IEEE Transactions on Affective Computing, <https://doi.org/10.1109/TAFFC.2020.3003816>
2. Benitez-Garcia G, Nakamura T, Kaneko M (2017) Methodical analysis of Western-Caucasian and East-Asian basic facial expressions of emotions based on specific facial regions. *Journal of Signal and Information Processing* 8:78–98
3. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E et al (2008) IEMOCAP: Interactive Emotional dyadic motion capture database. *Lang Resour Eval* 42(4):335–359
4. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (2011) Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Computer Vision Workshops, pp 2106–2112
5. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (2013) Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp 509–516
6. Dhall A, Kaur A, Goecke R, Gedeon T (2018) EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp 653–656
7. Douglas-Cowie E., Cowie R., Schroder M. (2000) A new emotion database: considerations, sources and scope, *proc. isca itrw on speech and emotion*, 3944
8. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. *J Pers Soc Psychol* 17(2):124–129
9. Hu P, Ramanan D (2017) Finding tiny faces. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1522–1530

10. Jack RE, Blais C, Scheepers C, Schyns PG, Caldara R (2009) Cultural confusions show that facial expressions are not universal. *Curr Biol* 19(18):1543–1548
11. Kolakowska A, Landowska A, Szwoch M, Szwoch W, Wróbel MR (2014) Emotion recognition and its applications. *Advances in Intelligent Systems and Computing* 300:51–62
12. Li S, Deng W (2018) Deep Facial Expression Recognition: A Survey. arXiv:1804.08348, Accessed 30 Sep 2019
13. Lim N (2016) Cultural differences in emotion: East-West differences in emotional arousal level. *Integrative Medicine Research* 5(2):105–109
14. Muszynski M., et al. (2019) Recognizing Induced Emotions of Movie Audiences From Multimodal Information. In: *IEEE Transactions on Affective Computing*, <https://doi.org/10.1109/TAFFC.2019.2902091>
15. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp 317–321
16. Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39(6):1161–1178
17. Simonyan K, Zisserman (2015) A Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556, Accessed 1 Sep 2019
18. Yu K, Wang Z, Zhuo L, Wang J, Chi Z, Feng D (2013) Learning realistic facial expressions from web images. *Pattern Recogn* 46(8):2144–2155
19. Zafeiriou S, Kollias D, Nicolaou M, Papaioannou A, Zhao G, Kotsia I (2017) Aff-wild: Valence and arousal in-the-wild challenge. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp 1980–1987
20. Zhang L, Tjondronegoro D, Chandran V (2014) Representation of facial expression categories in continuous arousal-valence space: Feature and correlation. *Image Vis Comput* 32(12):1067–1079
21. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23(10):1499–1503

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.