A Deep-Learning Based Model for Emotional Evaluation of Video Clips

Byoungjun Kim and Joonwhoan Lee

Division of Computer Science and Engineering, Chonbuk National University, Jeonju, Korea

Ijfis

Abstract

Emotional evaluation of video clips is the difficult task because it includes not only stationary objects as the background but also dynamic objects as the foreground. In addition, there are many video analysis problems to be solved beforehand to properly address the emotion-related tasks. Recently, however, the convolutional neural network (CNN)-based deep learning approach, opens the possibility by solving the action recognition problem. Inspired by the CNN-based action recognition technology, this paper challenges to evaluate the emotion of video clips. In the paper, we propose a deep learning model to capture the video features and evaluate the emotion of a video clip on Thayer 2D emotion space. In the model, the pre-trained convolutional 3D neural network (C3D) generates short-term spatiotemporal features of the video, LSTM accumulates those consecutive time-varying features to characterize long-term dynamic behaviors, and multilayer perceptron (MLP) evaluates emotion of a video clip by regression on the emotion space. Due to the limited number of labeled data, the C3D is employed to extract diverse spatiotemporal from various layers by transfer learning technique. The pre-trained C3D on the Sports-1M dataset and long short term memory (LSTM) followed by the MLP for regression are trained in end-to-end manner to fine-tune the C3D, and to adjust weights of LSTM and the MLP-type emotion estimator. The proposed method achieves the concordance correlation coefficient values of 0.6024 for valence and 0.6460 for arousal, respectively. We believe this emotional evaluation of video could be easily associated with appropriate music recommendation, once the music is emotionally evaluated in the same high-level emotional space.

Keywords: Video emotion analysis, C3D, Transfer learning, LSTM

Received: Dec. 2, 2018
 Revised : Dec. 15, 2018
 Accepted: Dec. 20, 2018

Correspondence to: Joonwhoan Lee
 (chlee@jbnu.ac.kr)
 ©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Automatic emotional evaluation of diverse media is a long-standing problem in multimedia analysis and artificial intelligence, which requires various component technologies, including feature extraction and selection from media, emotion analysis, and its evaluation. There have been many related types of research on automatic emotion evaluation. Yazdani et al. [1] has performed multimedia content analysis for emotional characterization of music video clips after extracting the audio and visual features. Also, Arifin and Cheung [2] has proposed the video content-based emotion analysis model for high-level video parsing problem. Zhao et al. [3] has proposed a video indexing and recommendation method based on emotion analysis focusing on the viewer's facial expression recognition. However, emotional evaluation of

video is still a difficult and challenging task because it includes not only stationary objects as the background but also dynamic objects as the foreground.

Recently, diverse convolutional neural network (CNN)-type deep learning techniques produce a big success in video analysis. However, traditional two-dimensional CNN (2DCNN) is not well suited for video because it handles only spatial information on a single image. Therefore, the recent researches on video analysis have been conducted with 3DCNN [4] and recurrent neural network (RNN) [5]. Unlike 2DCNN, the 3DCNN explores and exploits spatiotemporal features, even though the captured temporal feature is restricted within the short-time duration of the temporal window. Also, RNN can take a sequential input to characterize dynamic features and make the temporal decision associated with a relatively long-term behavior. Therefore, it has been successfully applied to various fields such as time series and video analysis.

In general, the most of the studies on deep learning-based video emotion analysis has been focused on the dynamic tracking of human facial expressions in video clips and the classification of them. The authors [6–8] have proposed deep learning-based facial image classification that uses the dynamic tracking of human facial expression in a video clip, where 2DCNN and RNN are combined in the structure. Also, Fan et al. [9] has proposed a facial emotion recognition for the combining 2DCNN-RNN and C3D. Vielzeuf et al. [10] has proposed a model of convolutional 3D neural network (C3D) and long short term memory (LSTM) for facial emotion recognition.

However, the structure combining 2DCNN and RNN for video analysis may not be sufficient to explore the information contained in video clips, because it extracts only spatial features from individual frames by 2DCNN, and shifts the responsibility to characterize the dynamic behavior to RNN. So the optical flow has been extracted to characterize the dynamic behavior separately to combine with the spatial information in video analysis [11]. The Vielzeuf's approach of C3D followed by LSTM, however, the C3D can extract both spatial and temporal behavior in a short time duration effectively, and succeeding LSTM can summarize the dynamic behavior to successfully perform facial emotion recognition.

In our conjecture, there are two problems when we directly employ the structure for facial emotion recognition to the evaluation of generic video emotion. One is the related scope of spatiotemporal features in C3D, and the other is the time duration that human can perceive the emotion.

In this paper, we propose a deep learning model for an emo-

tional evaluation of the video clips based on the conjectures. In the proposed structure, various short-term spatiotemporal features with different scopes are extracted from video clips using the C3D, and those features are aggregated to evaluate relatively long-term dynamic behavior through the LSTM [12].

Because a general type of video has diverse scenes with various shot speeds differently from the face, it is difficult to say that emotions depend only on high-level spatiotemporal features of the C3D. Rather the spatiotemporal features from various levels in the C3D could be useful to evaluate the general type of emotion of video clips. Note that the features from different levels in the C3D have their own scopes of 3D receptive fields in both spatial and temporal dimensions, where the features from lower layers have the smaller receptive fields than those from higher layers. In the proposed model, the emotion for a video clip is assumed to be dependent on the diverse spatiotemporal features from low to high levels of the C3D.

In general, video clips have the different length. So in the proposed model, the spatiotemporal features are temporally resized to be fixed by zero-padding before feeding into LSTM for temporal summarization. According to our conjecture, the summarization of the spatiotemporal features from C3D for a general emotion needs to take a longer duration than the facial expression recognition. The number of recursions in two-layer LSTM subnetwork reflects this conjecture in our model.

In the proposed model, the long-term dynamic features characterized by LSTM are mapped into Thayer 2D emotion space, which consists of valence and arousal dimensions, by the multi-layer perceptron (MLP)-type regression model. Therefore, the final emotion is estimated by the regression as a pair of valence and arousal scores in the emotion space.

In general, this type of data-driven machine learning approach requires a large amount of training data. Because the amount of available video emotion database contains only small labeled data, however, the proposed model uses adopts a transfer learning technique by taking the pre-trained C3D on the Sports-1M dataset of to capture the diverse spatiotemporal features. In the end-to-end training, the parameters of pre-trained C3D are fine-tuned with adjusting the weights of LSTM and regression network simultaneously. After training the video emotion one can feel when experiencing a video clip is evaluated in the 2D emotion space.

The experimental results show that the concordance correlation coefficient values of 0.6024 for valence and 0.6460 for arousal, respectively. This shows that the deliberate construction of a deep learning structure, C3D and then LSTMs followed

by the regression network, can provide excellent performance to obtain the emotion, which is one of the high-level video semantics. Therefore, the contribution of this paper can be described as follows.

First, we propose a deep learning based emotion evaluation model of a general type of video clips, which consists of C3D, and then LSTMs followed by regression network.

Second, the proposed model is constructed based on two conjectures, one is the emotion can be well captured by various levels of spatiotemporal features from C3D, and the other is the general emotion from a video clip needs to summarize dynamic features for a longer time than facial expression in RNN.

Third, the proposed model can produce excellent performance compared with other methods.

The paper is organized as follows: the related works are reviewed in Section 2, the details of our proposed method are given in Section 3, and the experimental results and analysis are presented in Section 4. Finally, concluding remarks are given in Section 5.

2. Related Works

2.1 Emotion Space

In general, the emotion of multimedia contents that human feel when they are experienced can be expressed with adjectives. But there are too many adjective terms to be used in the expression. For this reason, psychological experiment with statistical analysis has been done to reduce the number of vocabulary and to define a compressed emotion space, on which adjectives are distinguished as a position. For example, Kawamoto and Soen [13] has analyzed emotions of color patterns with 13 adjective pairs, and Lee and Park [14] has defined three pairs of adjectives consisting of “warm-cold”, “heavy-light” and “dynamic-static” for emotional evaluation of color images. Another well-known 2D emotion space defined by Thayer [15] is the valence-arousal space. In the space, the valence dimension refers to the positive/negative intrinsic attractiveness, and the arousal dimension expresses the degree of being state of awoken or of sense organs stimulated to a point of perception. According to the amounts of valence and arousal, 12 distinctive adjectives can be located on the boundary of a unit circle as in Figure 1. This paper uses this continuous Thayer’s emotion space to evaluate video and the result is given by a position on the unit circle in 2D space.

Figure 1. Thayer’s 2D emotion space [15].

2.2 Video Analysis with CNN-Based Spatiotemporal Features

In earlier research of CNN-based video analysis, transfer learning has been used [16], where features are independently extracted from the individual frame of video. However, there is a problem in video analysis, because it ignores the temporal information involved in a video clip because only the spatial features are considered. Donahue et al. [17] has proposed a network model combining 2DCNN and LSTM for the video analysis. This method extracts spatial features sequentially from consecutive frames using 2DCNN and then shifts the responsibility to capture the temporal information to LSTM. But this structure provides a limited performance because the 2DCNN extracts only spatial features and the LSTM may have too much burden to extract temporal features and characterize the dynamic features simultaneously. In addition, the 2DCNN with LSTM structure usually extracts the most abstract spatial features from the last fully connected layer following the consecutive 2DCNNs, and feeds them into LSTM. They do not utilize the low-level spatial features from the lower CNN layers so that there might occur information loss.

Simonyan and Zisserman [11] have shown a high performance on video analysis with a two-stream network that consists of two branches of 2DCNNs, one for capturing spatial features and another for temporal behavior from optical flow. But the two-stream network requires a large number of computations because the optical flow must be created in the preprocessing process.

Recently, the focus of video analysis has changed according

Figure 2. Differences between 2D and 3D convolution operations [18]. (a) A 2D feature map with spatial features as output after performing 2D convolution kernel. (b) A feature map with spatiotemporal features as volume output after performing 3D convolution kernel.

to the emergence of 3DCNN. A 3DCNN has provided good results in video analysis because it captures spatiotemporal features by 3D convolution kernels and pooling operations. Figure 2 shows that the difference between 2D and 3D convolution operations. Note the scope of the temporal features in 3DCNN is restricted because the number of layers should be increased to capture the long-term temporal features that make the training difficult. Therefore, this 3DCNN captures temporal features defined only in a short-time duration, and not appropriate to characterize a long-term dynamic behavior.

2.3 The C3D Structure

In general, 3D convolution receives $W \times W \times F \times K$ image data as input. Here, $W \times W$ denotes the size of the image, F is the number of frames to be processed at once, and K is the number of channels, respectively. In addition, the filter used in the 3D convolution has a size of $H \times H \times R \times K$, $H \times H$ is the horizontal and vertical size of the filter, R is the number of frames, and K is the number of channels, respectively. Like 2D convolution, the 3D filter takes a convolution operation by moving as much as a stride in horizontal and vertical directions, and across the time axis. This can be expressed as

$$v_{ij}^{xyz} = \tanh \left(b_{ij} \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)jm}^{(x+p)(y+q)(z+r)} \right), \quad (1)$$

Figure 3. The C3D structure [18].

where (x, y, z) is the volume coordinates of a feature map, (p, q, r) is the spatial and time dimension kernel index, and j/m is the feature map/volume index, and i is the convolutional layer index, respectively.

The C3D can model appearance and motion information simultaneously and outperforms the 2DCNN features on various video analysis tasks. In addition, the deeper architecture with a uniform $3 \times 3 \times 3$ kernel size has been empirically verified to produce the best result [18]. The C3D input dimension is $112 \times 112 \times 3 \times 16$, consisting of 5 groups of convolutional layers, 5 pooling layers, 2 fully connected layers and softmax. The numbers of filters in the convolution layers are 64, 128, 256, 512, and 512, respectively. There is no size change between convolution layers, that is $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$ and proper zero-padding. All pooling layers used max pooling of kernel size $2 \times 2 \times 2$ except pool1 to preserve the temporal information in the early phase. Figure 3 shows that the C3D structure.

3. Proposed Method

In this section, the proposed emotion evaluation model is explained in detail, including how to extract the spatiotemporal features of video clips using C3D and how to evaluate the emotion using LSTM with MLP-type regression network.

3.1 The Proposed Emotion Evaluation Model

The learning ability of a deep learning model is related to the network depth. In general, even though a deeper network increases the expressiveness, it also makes the training and optimization difficult. In addition, the deeper network with a large number of parameters requires a large amount of training data to avoid overfitting problem. In the proposed structure, various short-term spatiotemporal features are extracted from a video clip using C3D based on transfer learning technique, and those features are aggregated to characterize the long-term dynamic behavior through LSTM. In the evaluation process, emotion is estimated by regression mapping realized by the MLP to valence and arousal axes on Thayer 2D emotion space. Figure 4 shows the proposed structure of emotional evaluation model.

Figure 4. The proposed emotion evaluation model.

3.2 The Spatiotemporal Feature Extraction

The C3D can act as a feature extractor for video analysis [18]. A video clip passing to the C3D produces spatiotemporal features through the successive convolutional layers and corresponding max-pooling layers. In video clips, it is difficult to say that emotions depend only on the highest-level spatiotemporal features of the C3D. Rather the spatiotemporal features from various levels in C3D layers could be useful to evaluate an emotion from the general type of video clips.

In the proposed method, each video clips passing to the C3D extract the 960 features from the 1st to the 4th max-pooling layers through GAP (global average pooling) and concatenate them with the 4096-dimensional feature vectors in the first fully connected layer. Note that the GAP results of the 1st thru the 4th layer represent the globally averaged spatiotemporal features of the smallest and the largest ranges of receptive fields, respectively. Also, the outputs from the fully connected layer can be interpreted as the most highest-level features for generic action recognition. Therefore, the spatiotemporal features in the proposed model include all the globally aggregated ones with diverse spatiotemporal scopes. Note that the local spatiotemporal features are available unless we take GAP, but it could increase the number of features to make a burden to the following stage of LSTM. This is the reason why we adopt the GAP to aggregate the spatiotemporal features at each level of C3D. Figure 5 shows the low-level and high-level spatiotemporal feature extraction.

3.3 LSTM with MLP-Type Regression Network

In recent study, LSTM has been used to estimate the emotion of a video clip in the same Thayer's emotion space. But the hand-crafted audio and video features were extracted, then only the selected features were exploited to estimate the degrees of arousal and valence in the LSTM [19]. In the work, the

Figure 5. The proposed spatiotemporal feature extraction.

Figure 6. The LSTM with MLP-type regression structure.

idea of LSTM is similar to ours, in that it takes a role to characterize a long-term dynamic behavior of video clips. In our proposed model, however, LSTM is adjusted concurrently with the fine-tuning of C3D to automatically define the spatiotemporal features in end-to-end training. Note that the duration of a set of C3D features is only 16 frames, so that the temporal components is not enough to define the complicate long-term emotions unless the original video frames are subsampled. That's why LSTM is necessary to be added for summarizing the long-term temporal behavior.

In the proposed method, we evaluate the emotion of the extracted spatiotemporal features using LSTM with MLP-type regression network. The proposed structure consists of two LSTM networks and a two-layer MLP network. There are 1024 internal states in each of the two LSTMs. The input of the MLP-type regression network is equal to the number of states in the second LSTM, the number of hidden units is 256, and there are two output units with tangent hyperbolic activation functions which correspond to estimated scores on arousal and valence dimensions. Figure 6 shows the LSTM with MLP-type regression network.

Figure 7. Screenshot samples of video clips and their corresponding annotated valence (V) and arousal (A) values.

4. Experimental Results

4.1 Data Collection and Spatiotemporal Feature Extraction

To evaluate the performance of the proposed model, a dataset with 12,900 videos clips is constructed by MediaEval 2015 Affective Impact of Movie task [20], and selected YouTube; about 9,800 video clips from MediaEval 2015 and the rest from our own selection of YouTube. The MediaEval data includes movies with proper labels of arousal and valence values on $\{-1, 0, 1\}$, while the selected YouTube consists of dramas, sports, and so on. The criterion for video selection from YouTube is that the content is supposed to elicit strong emotional reactions to viewers such as happiness, excitement, fear, anger, and so on. According to the adjectives, video clips whose lengths are from 7 to 18 seconds have been collected. Then we have annotated scores on the valence and arousal dimensions by one of five levels, i.e., $[-1, -0.5, 0, 0.5, 1]$. Take the valence dimension for instance, -1 indicates that there is negative emotion in the video content, while 1 stands for the positive emotion. Figure 7 shows screenshot examples of the video clips and their annotations.

Among 12,900, we have randomly selected 11,000 video clips for training, and the remaining 1,400 for validation, and 500 for testing. Then the pre-trained weights C3D from benchmark Sports-1M dataset are used to initialize the C3D, we extracted the spatiotemporal features for video clip using the proposed feature extraction method.

Figure 8. The combining LSTM with MLP-type regression network.

4.2 Combining LSTM with MLP-Type Regression Network

The connection of C3D and LSTM networks is shown in Figure 8, where every set of successive 16 frames overlapped with 8 frames produces a set of spatiotemporal features to feed into LSTM network, which means the size of the temporal window is 16 frames with the stride of 8 frames. The LSTM network produces the summarized dynamic behavior over 60 temporal windows, which implies an emotion is evaluated for every sequence of 488 frames. Because the maximum number of windows, which corresponds to the number of recursions in LSTM network, is assumed to be 60, the clips with the smaller number of windows than 60 perform zero-padding.

The captured dynamic features of time-varying features for consecutive 488 frames are fed into regression network in order to estimate the emotion.

For the experiment, Intel Core i5-6600 CPU with the GTX 1080 Ti GPU has been used. The proposed model is implemented with TensorFlow library. The loss function to minimize for end-to-end learning is the mean squared error between the estimated emotion and ground truth of the video clips. Also, to avoid overfitting, l_2 regularization is included in the cost function.

In addition, we used dropout when the first and second layers of LSTM results are loaded to the second layer of LSTM*¹ and MLP-type regression network*², respectively. The training method is summarized in Table 1.

4.3 Experimental Results and Discussion

For evaluation metrics, we choose the CCC, the Pearson correlation coefficient (PCC), and R²-score. CCC is a measure of agreement between two variables on $[-1, 1]$. PCC measures the strength of a linear relationship between two variables on $[-1, 1]$, where 1, 0, and -1 means a positive, no, and negative correlations, respectively. R²-score is a measure for the

Table 1. The training methods

Method	Parameter
Batch size	50
Number of epochs	500
Dropout	0.7^{*1} , 0.5^{*2}
Weight initialization in LSTM/MLP regression network	Xavier
Weight decay	0.0005
Learning rate	$1e-4$
Optimizer	RMSProp

Table 2. Emotional evaluation results of the video clip

	Valence			Arousal		
	CCC	PCC	R^2	CCC	PCC	R^2
High-level only	0.5381	0.5464	0.2986	0.5682	0.5701	0.3251
Proposed method	0.6024	0.6049	0.3661	0.6460	0.6471	0.4188

Table 3. Emotional evaluation of video clips for the LSTM time step

LSTM time step	Valence			Arousal		
	CCC	PCC	R^2	CCC	PCC	R^2
20	0.3078	0.3251	0.1124	0.2671	0.2876	0.1754
60	0.6024	0.6049	0.3661	0.6460	0.6471	0.4188

predicted fitness on $[0, 1]$ for a regression network.

Table 2 shows the results of the emotion evaluation with the proposed model when the highest level with or without the lower level spatiotemporal features from C3D are used as input to the LSTM network. As we expected, the highest level with the lower level features as in the proposed method provides the better results. This implies not only the highest-level features for action recognition but also the lower level spatiotemporal features are helpful to improve the emotional evaluation of a video clip.

Figure 9 shows the evaluation results of valence and arousal scores in the Thayer's emotional space. The proposed model produces the similar evaluation scores as the human can feel on the video clips. Note that the video clips are well aligned with the adjectives, for example, nervous video clips can be easily distinguished from pleasant ones.

Table 3 shows performance comparisons of the proposed model with different time steps of LSTM, which correspond

Figure 9. The Thayer's emotional space based on the emotional evaluation results of the proposed method.

to the length of video clips for emotion evaluation. Note that the proper length of video clips is important to the better performance of emotion evaluation. Evidently, 20 frames are not enough to evaluate proper emotion from video clips similar to human being. Note that the 20 and 60 frames correspond to $2/3$ and 2 seconds, respectively, when the frame rate is assumed to be 30/sec.

Table 4 shows that the CCC values of different methods. Because the video clips for the emotional evaluation are different to each other and our method uses only the visual features without audio of a video clip, the direct comparison is not possible. But our method gives CCC values of 0.6024 for valence and 0.6460 for arousal, respectively. Note that our proposed method provides better performance than the methods of Baveye et al. [20] and Gan et al. [21], even though we use only visual information. Also, our method shows comparable or better performance with the method of Zhang and Zhang [19] where it uses carefully designed and selected handcrafted features. The results show the proposed model is promising in the emotional evaluation of video clips.

5. Conclusion

In this paper, we have proposed a deep learning model for an emotional evaluation of a video clip. In the proposed model, the pre-trained C3D network generates short-term spatiotemporal video features in the various levels, LSTM network accumulates the consecutive time-varying features to characterize long-term dynamic behaviors, and MLP network evaluates emotion of a video clip by regression mapping on the Thayer's 2D emotion

Table 4. Performance comparisons of the emotional evaluation of video clip

Method	CCC	
	Valence	Arousal
Baveye et al. [20], with audio visual information	0.542	0.645
Gan et al. [21], with audio visual information	0.398	0.429
Zhang and Zhang [19] with only video information	0.645	0.542
Proposed method	0.602	0.646

space.

Due to the limited number of labeled data, the C3D features of diverse levels are extracted by the transfer learning technique. The pre-trained C3D on the Sports-1M dataset and LSTM network followed by the MLP for regression mapping are trained in end-to-end manner to fine-tune the C3D, and to adjust weights of two LSTMs and the MLP-type emotion estimator.

For the training of the model and the performance evaluation, MediaEval 2015 and our own selected and annotated YouTube data have been used. Through the experiments, we have empirically shown that both the low-level and high-level spatiotemporal features are useful to improve the accuracy of emotion evaluation on the valence and arousal dimensions, and proper LSTM network is useful to characterize the long-term temporal dynamic behavior like emotion. The proposed method achieves the CCC values of 0.6024 for valence and 0.6460 for arousal, respectively. We believe this emotional evaluation of video could be easily associated with appropriate music recommendation, once the music is emotionally evaluated in the same high-level emotional space. Our future work will be conducted in two directions. The one is to find more effective structures of deep models for audio and video spatiotemporal features extraction. The other is to develop an efficient method for multimodal emotion evaluation.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgement

The research leading to these result, authors would like to thank Korean Ministry of Education for funding. The National Research Foundation of Korea (NRF) support us under Basic

Science Research Program (NRF-2015R1D1A1A01058062).

References

- [1] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, article no. 26, 2013. <https://doi.org/10.1186/1687-5281-2013-26>
- [2] S. Arifin and P. Y. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325-1341, 2008. <https://doi.org/10.1109/TMM.2008.2004911>
- [3] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, AZ, 2011, pp. 1473-1476. <https://doi.org/10.1145/2072298.2072043>
- [4] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2013. <https://doi.org/10.1109/TPAMI.2012.59>
- [5] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," 2017. Available <https://arxiv.org/abs/1801.01078>
- [6] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *Proceedings of 2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 619-623. <https://doi.org/10.1109/ICIP.2016.7532431>
- [7] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101-106, 2018. <https://doi.org/10.1016/j.patrec.2018.04.010>
- [8] D. Kollias and S. Zafeiriou, "A Multi-component CNN-RNN Approach for Dimensional Emotion Recognition in-the-wild," 2018. Available <https://arxiv.org/abs/1805.01452>

- [9] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016, pp. 445-450. <https://doi.org/10.1145/2993148.2997632>
- [10] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, UK, 2017, pp. 569-576. <https://doi.org/10.1145/3136755.3143011>
- [11] K. Simonyan and A. Zisserman, A. Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 27, pp. 568-576, 2014.
- [12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [13] N. Kawamoto and T. Soen, "Objective evaluation of color design. II," *Color Research & Application*, vol. 18, no. 4, pp. 260-266, 1993. <https://doi.org/10.1002/col.5080180409>
- [14] J. Lee and E. Park, "Fuzzy similarity-based emotional classification of color images," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1031-1039, 2011. <https://doi.org/10.1109/TMM.2011.2158530>
- [15] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York, NY: Oxford University Press, 1989.
- [16] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4724-4733. <https://doi.org/10.1109/CVPR.2017.502>
- [17] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 2625-2634. <https://doi.org/10.1109/CVPR.2015.7298878>
- [18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 4489-4497. <https://doi.org/10.1109/ICCV.2015.510>
- [19] L. Zhang and J. Zhang, "Synchronous prediction of arousal and valence using LSTM network for affective video content analysis," in *Proceedings of 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Guilin, China, 2017, pp. 727-732. <https://doi.org/10.1109/FSKD.2017.8393364>
- [20] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: a video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43-55, 2015. <https://doi.org/10.1109/TAFFC.2015.2396531>
- [21] Q. Gan, S. Wang, L. Hao, and Q. Ji, "A multimodal deep regression bayesian network for affective video content analyses," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 5113-5122. <https://doi.org/10.1109/ICCV.2017.547>

Byoungjun Kim received his B.S. and M.S. degrees in division of computer science and engineering from the Chonbuk National University, Korea, in 2015 and 2017, respectively. He is currently a Ph.D. fellow in Chonbuk National University. His research interests are computer vision, deep learning, and artificial intelligence.

E-mail: breed213@jbnu.ac.kr

Joonwhoan Lee received his B.S. degree in Electronic Engineering from the Hanyang University, Korea in 1980. He received his M.S. degree in Electrical and Electronics Engineering from KAIST, Korea in 1982, and the Ph.D. degree in Electrical and Computer Engineering from University of Missouri, USA in 1990. He is currently a Professor in Division of Computer Science and Engineering from the Chonbuk National University, Korea. His research interests include image and audio processing, computer vision, and artificial intelligence.

E-mail: chlee@jbnu.ac.kr