https://doi.org/10.5392/IJoC.2019.15.4.059

Gesture-Based Emotion Recognition by 3D-CNN and LSTM with Keyframes Selection

Son Thai Ly, Guee-Sang Lee, Soo-Hyung Kim, Hyung-Jeong Yang
Department of Electronics and Computer Engineering
Chonnam National University, Gwangju, South Korea, 61186

ABSTRACT

In recent years, emotion recognition has been an interesting and challenging topic. Compared to facial expressions and speech modality, gesture-based emotion recognition has not received much attention with only a few efforts using traditional hand-crafted methods. These approaches require major computational costs and do not offer many opportunities for improvement as most of the science community is conducting their research based on the deep learning technique. In this paper, we propose an end-to-end deep learning approach for classifying emotions based on bodily gestures. In particular, the informative keyframes are first extracted from raw videos as input for the 3D-CNN deep network. The 3D-CNN exploits the short-term spatiotemporal information of gesture features from selected keyframes, and the convolutional LSTM networks learn the long-term feature from the features results of 3D-CNN. The experimental results on the FABO dataset exceed most of the traditional methods results and achieve state-of-the-art results for the deep learning-based technique for gesture-based emotion recognition.

***Key words**: Gesture-based Emotion Recognition, 3D Convolutional Networks, Convolution LSTM.*

1. INTRODUCTION

In recent years, emotion recognition has been a growing interest in psychology and computer science community. This is due to the wide range of applications, including intelligent human-computer interaction, healthcare, entertainment industries, etc.

In the task of emotion recognition, there are three main modalities that have recently focused on: facial expression, speech, and bodily gesture for its potential of analyzing affective information. Body expression could be considered, as all the motions of the body, namely, gesture, body postures, and movement. Although, compared to other modalities, bodily expression delivers a majority portion of important information in recognizing emotion [1], yet it has not drawn much attention from the science community. Generally, all the affective gesture recognition research must carry based on the results of psychological research which indicated the effect of gesture. For a comprehensive understanding on emotion recognition via the body gesture, the works of Kleinsmith and Bianchi-Berthouze [1], Karg et al. [2] and Glowinski et al. [3], [4] should be referred. Although there are several studies about gesture-based emotion recognition, most of them heavily used the hand-crafted features for the task, leading to significant computational cost, such as the works in [2]-[11].

Although psychological researches reveal that bodily gestures convey crucial information for emotion recognition problems, facial expression and speech analysis have been dominant and used in a vast majority of a number of studies. Still, there are studies about gesture-based emotion recognition and most of them heavily used the hand-crafted features. For example, Gunes and Piccardi work [7], [8] applied several classifiers, namely, Random Forest, BayesNet, Adaboost, SVM for emotion classification. Chen's research [9] considers the entire the expression cycle, i.e., neutral, onset, apex, offset by exploiting the temporal normalization method. Hence, the approach could better capture the dynamic information of the expression than apex frame. Generally, these traditional hand-crafted features methods are not considered as optimal methods for emotion recognition. It also does not have much room for improvement as most of the science community are carrying their research by deep learning technique.

Compared to above methods, the neural network is more considerable as human brain model. Especially, the convolutional neural networks (CNN) were first introduced by Lecun et al. [12]. They are easily customized in a various number of layers with different size of filters, and the network creates a representation of each individual input and integrates them in the last layer. Due to the various visual representation these models could achieve, they are applied in several tasks, such as object detection, and tracking. However, their model poorly performs at sequences level which is crucial for emotion recognition via gesture. To cope with this weakness, Barros et al. [13] made use of 3D-CNN [14] to deal with the sequence of frames. Although they achieved notable results, the authors

** Corresponding author, Email: gslee@jnu.ac.kr*

Manuscript received Jul. 05, 2019; revised Oct. 02, 2019; accepted Oct. 08, 2019

employed the temporal phase which is annotated in the dataset. Another approach is presented in Ranganathan et al. work [15], they recruited actors and constructed a specific dataset to train face-body-voice multimodal convolutional deep belief model. Despite the competitive result, their work is time and resource-consuming, also difficult to adapt.

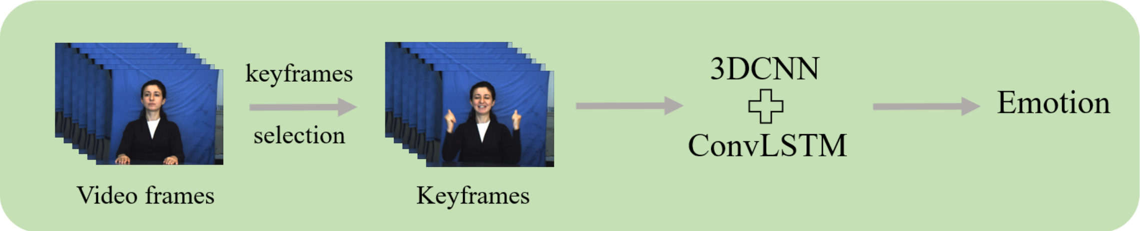

Fig. 1. The pipeline of our framework. Firstly, keyframes are extracted from raw video data. It then is considered as input for the 3D CNN coupled with ConvLSTM network emotion classification

To cope with these drawbacks, we present a deep learning model to classify the emotion from raw videos as input and emotion label as outputs. Firstly, the keyframes are extract the keyframes from raw video. The extracted keyframes are then used as input of 3D convolutional neural networks (3D CNN) coupled with convolutional LSTM network (ConvLSTM) networks to explore both short-term and long-term spatiotemporal gesture features. The overview of the proposed model is shown in Fig. 1. The model is evaluated on FABO dataset [7] and achieved the state-of-the-art result on gesture-based emotion recognition. The results also indicate that ConvLSTM effectively reduces the computational cost and, hence the processing time. Furthermore, our model is end-to-end training model. It means we do everything automatically from selecting keyframes to classifying the emotion. It is worth mentioning that although the dataset also includes the temporal annotation for each video which indicates the informative frames, unlike [13], we neglect it.

In summary, our contributions are in three-fold as follows:

1. The idea of frame selection has never been applied before in the literature.
2. An end-to-end framework for gesture-based emotion recognition from raw video.
3. Deep architectures for gesture-based emotion recognition based on 3D convolution and convolution LSTM which are originally developed for action recognition.
4. The state-of-the-art results on FABO dataset for using end-to-end deep learning method.

The remaining sections are organized as follows: Section 2 mentions the methodologies of this study. Section 3 describes the experiments, results along with the discussions. The conclusion of our approach and further works are stated in Section 4.

The preliminary version of this paper was presented in [16]. The major extensions added to this paper are as followed: (1) The ITQ keyframes selection method is described in more detail in Section 2.1; (2) The experiments on effect of different keyframe retrieving methods are added in Section 3.3.

2. PROPOSED METHOD

2.1 Keyframe selection

To boost the classifier's performance, we need to retrieve keyframes from raw sequences. Keyframe help the network to distinguish one sequence from others much easier. Therefore, selecting keyframes from video is crucial. In this work, we compare three keyframe retrieving methods: overlapping windows, k-mean clustering, and iterative quantization.

Overlapping windows is the method that collects keyframes by skipping a fixed number of frames in the video. Gestures that describe emotion are usually hastily carried out in 1-2 second. Hence, this method is neither efficient nor optimized. Recently, computer vision tasks have related to the problem of learning binary codes for representing the images and large-scale image retrieval. Torralba et al. [17] raised the binary coding as a computer vision problem and compared several methods, such as boosting, Boltzmann machines [18], and LSH [19]. Generally, in binary coding, PCA must be performed as a common initial step to reduce the dimensionality of the data. Hence, the variance between directions is different, leading to the problems of poor performance when binarizing the same number of bits for every direction. There are many studies proposing solutions for this problem, for example, Spectral Hashing [20] using Laplacian eigenfunction formulation or Shift-invariant kernels [21]. The iterative quantization method used in this work is employed from Gong et al. work [22] for its simplification, effectiveness, and easily implemented for selecting the keyframe. The data dimension is first reduced by performing PCA, and then applied the random orthogonal transformation to balancing the variance between eigenvectors' directions. The transformation is then iterated to minimize the quantization error. We employed the ITQ method coupled with PCA to hash each frame of video into fixed 16-bits binary code. First, each frame is extracted into features in length 4096 by pre-trained VGG16 [23]. We utilize the *fc7*'s output of this pre-trained network to extract the spatial representation of the frame. Since the feature must be zero-centered in order to binarize by ITQ, we then applied PCA. After that, ITQ was performed for 50 iterations as suggested in [22]. The keyframes are then selected by considering the Hamming distance between two consecutive frames. The iterative quantization keyframes retrieval method is illustrated in Figure 2.

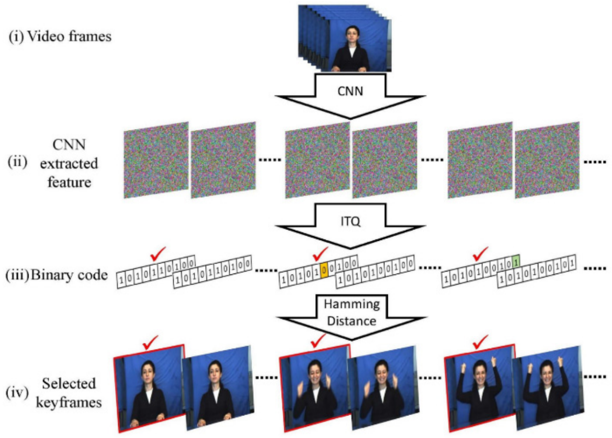

Fig. 2. Outline of the iterative quantization keyframe selection procedure. First, features are extracted from every raw frame by pre-trained VGG16. The feature maps are then binarized by ITQ method and, finally, the Hamming distance is applied to retrieve the keyframes

2.2 3D Convolutional Networks

3D-CNN [24], is an effective 3D convolutional neural networks for spatiotemporal feature learning. Unlike 2D convolutional networks which could learn only the spatial features for later learn the temporal features by the LSTM, the 3D ConvNets is capable of processing spatiotemporal information in one shot. This property of 3D-CNN is proved especially effective for action recognition in many studies in this topic as in [25].

2.3 Convolutional LSTM Networks

For expression emotion in gesture, there is no set of pre-determined motion. Moreover, the motion pattern is usually spontaneous and distinct for each person. This leads to the memory problems, giving when to "forget" or "update" new information. The recurrent neural network is usually preferred for dealing with this type of problems.

2.3 Convolutional LSTM Networks

For expression emotion in gesture, there is no set of pre-determined motion. Moreover, the motion pattern is usually spontaneous and distinct for each person. This leads to the memory problems, giving when to "forget" or "update" new information. The recurrent neural network is usually preferred for dealing with this type of problems.

In this work, we employ the ConvLSTM [26], a variant of LSTM. ConvLSTM has small memory usage by using the convolutional operation, instead of the fully connection in LSTM. By using the convolution operation, the ConvLSTM could learn the representation of spatiotemporal correlation better. In other words, it improves the spatiotemporal smoothness between local pixels, compared to the fully connected layer in LSTM. Another benefit of using convolution operation in ConvLSTM is that it has smaller memory usage, leading to reduce the network's parameters. The utilized ConvLSTM cells in this paper are similar as in [25]:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \\
f_t &= \sigma\left(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f\right) \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \\
g_t &= tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes g_t \\
h_t &= o_t \otimes tanh(c_t)
\end{aligned}
$$

Where, *, ⊗, and σ denote convolution operation, element-wise product, and the sigmoid function respectively. $W_{xi}, W_{hi}, W_{xf}, W_{hf}, W_{xo}, W_{ho}, W_{xc}, W_{hc}$ are 2D Convolutional kernels. b_i, b_f, b_o, b_c are the biases. While i_t is the input gate, $f_t, o_t,$ and g_t are the forget gate, input gate and input modulation gate, respectively. The deep network composed of 3D CNN coupled with ConvLSTM is shown in Fig. 3.

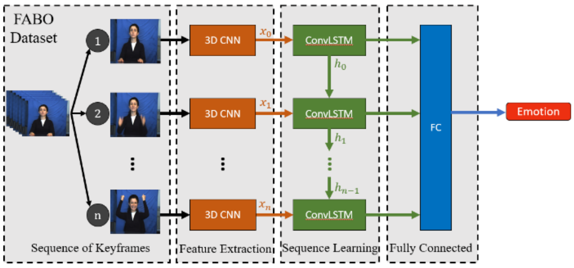

Fig. 3. Deep architecture 3D CNN coupled with ConvLSTM for gesture-based emotion recognition.

2.3 Grayscale overlapping

As an addition learned features, we reserve the sequence of all extracted keyframes by stacking into one grayscale image. This stacked image is then considered as input for typical 2D CNN to classify the emotion. This technique is replicated as in Barros et al. study [13] in which stacked only four temporal phase frames: onset, apex, offset and neutral. Slightly different from their work, in this study, we stack all the extracted keyframes into one frame as mathematically expressed as:

$$M = \sum_{i=1}^{N} |(F_{i-1} - F_i)|(i/t)$$

Where resulting image M is the sum of absolute difference between N consecutive frames in the sequence at a specific gray scale with t increasing overtime until the end of the sequence. The absolute difference also removes the static background. The example result is shown in Fig. 4.

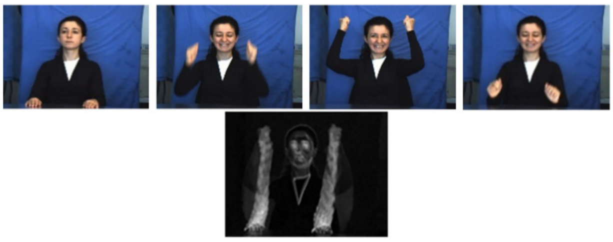

Fig. 4. Example of a grayscale overlapping image. Here, the example only shows 4 images while it needs the entire sequence to form the result image

3. EXPERIMENTS AND RESULTS

3.1 Datasets and Experiment Environment

We used the FABO dataset[7] to evaluate our framework. The dataset contains 281 videos in which express ten emotional states: ‘‘Anger’’, ‘‘Anxiety’’, ‘‘Boredom’’, ‘‘Disgust’’, ‘‘Fear’’, ‘‘Happiness’’, ‘‘Surprise’’, ‘‘Puzzlement’’, ‘‘Sadness’’ and ‘‘Uncertainty’’. The FABO dataset was commonly used for emotion recognition with gestures until recently [28] and it showed that the combination of facial features and body movements significantly improve the detection of emotion relevant areas in the image. This corpus contains recordings of the upper torso of 23 subjects. As mentioned before, we do not use the temporal annotation provided by the dataset to manually get the keyframes as in [13]. All experiments were performed 10-fold cross-validation on the dataset. The experiment was accomplished with Intel i7-6700 CPU and Nvidia GTX 1080 GPU.

3.2 Input Processing and Baselines

In this study, given an arbitrary length video, a set of *n* keyframes was retrieved and formed as an input sequence. For the experiments, the set of *n* input keyframes was resized into n×128×128×3. Here, we experimentally set n = 16.

We created a baseline to evaluate the effect of the 3D CNN coupled ConvLSTM, compared to randomly sampled frames and iterative quantization extracted keyframes which both were processed by Long-term Recurrent Convolutional Networks [27] (LRCN) and 3D-CNN coupled ConvLSTM, respectively. In addition, we also examined the effect of different keyframe retrieving methods: overlapping windows, k-mean clustering, and iterative quantization. Unless stated otherwise, the experiments were conducted with 16-keyframe extracted by learned by iterative quantization method and learned by 3D CNN coupled ConvLSTM.

3.3 Results

3.3.1 Baseline results: The results of baselines’ accuracy are shown in Table 1. The accuracy of 3D CNN and ConvLSTM is highest with 66.6%, while 40-random-frame and 16-keyframe processed by LRCN are only 53.3% and 63.3%, respectively. It suggests that the keyframes extraction has effect on emotion recognition. Furthermore, the 3D CNN ConvLSTM exceeds 3% over the LRCN, indicating the advantage of better representation learning from convolution operation within ConvLSTM. 3D CNN are particularly good for learning the spatiotemporal feature simultaneously.

Table 1. Comparison of the model and baselines

Experiments	Accuracy
40-random-frame LRCN	53.3%
16-keyframe LRCN	63.3%
16-keyframe 3D CNN coupled ConvLSTM	**66.6%**

3.3.2 Effect of different keyframe retrieving methods: Table 2 report the results of three keyframe selection methods: overlapping windows, k-mean clustering, and iterative quantization. We can see that the iterative quantization has the highest performance.

Table 2. Effect of different keyframe retrieving methods

Experiments	Accuracy
Overlapping windows	56.6%
k-mean clustering	60%
Iterative quantization	**66.6%**

3.3.3 Comparison to state-of-the-art: We also compare our proposed framework with other studies on same FABO dataset as shown in Table II. Our framework has the higher accuracy than most of the non-deep-learning-based methods in [8], [9], and multi-channel CNN deep learning methods [13], except for random forest in [8].

The work in [8] was composed of optical flow, edginess, geometry features representation among others. As a consequence, it needs accurate motion modelling which can be difficult with different environments, e.g. images with illumination changes or background noises.

Each body part, namely head, hands, shoulder, was extracted into 144 different representation features. As a result, these hand-crafted features are likely to require more computation than our method. Compared to results in [8], our 3D CNN coupled ConvLSTM result is better than SVM 2% and worse than random forest 10%. The reason for this could be that the model needs more structured motion representation. Also this random forest technique is effective when the dataset is small. When large dataset is available, deep learning framework can be more effective in general. This paper has presented the deep learning framework which turned out to be promising and can be further improved. Compared to other existing deep learning based work [13], our result generated a superb result.

In [9], the Histogram of Oriented Gradient, Motion History Image, and skin color segmentation were used hands, shoulders, and face position tracking. Although the features learning was heavily relied on tracking body components, the results in both bag of words and temporal normalization methods are about same as 3D CNN coupled ConvLSTM our framework; however, our ensemble result is 7% higher. It is also worth mentioning that our framework took only around 7 minutes for training.

To the deep learning based method, accuracy of our 3D CNN ConvLSTM result is over 8% better than multichannel convolutional neural network (MCCNN) [13]. This could be due to the fact that we benefit from not only the entire selected keyframes sequence as input but also the spatiotemporal smoothness in local pixel by the convolution operation within ConvLSTM; whereas in [13], only manual selected 4 frames which were then feature extracted by 3D CNN, were chosen from the sequence. In addition, the grayscale overlapping method is a simple, yet effective with 56.6% of accuracy when compared to other complicated methods. This prove the positive effect of the keyframes selection could increase the overall performance. Moreover, our framework employed the keyframes selection which is more appropriate for real application in the future.

Another thing to be noticed is that the number of videos in FABO dataset is small for deep learning method. This also holds true for [13].

Table 3. Comparison of the state-of-the-art for gesture-based emotion recognition on FABO dataset

Approach	Accuracy
SVM [8]	64.5%
Random forest [8]	**76%**
Temporal normalization [9]	66.7%
Bag of words [9]	65.3%
Grayscale overlapping [13]	53.32%
MCCNN [13]	57.84%
Our grayscale overlapping	56.6%
3D CNN coupled ConvLSTM	66.6%
Ensemble our two models	73.33%

4. CONCLUSION AND FUTHER WORK

In this study, we introduced a deep learning framework for emotion gesture recognition and evaluated it on FABO dataset. The results point out that our end-to-end deep learning framework, composed of keyframe extraction, and 3D CNN coupled with ConvLSTM, outperforms other hand-crafted feature methods.

It suggests the potential of enhancing the performance and it also indicates limitation in terms of well-rounded dataset for gesture emotion recognition. Therefore, in future work, we intend to reuse other emotion datasets and setup the gesture-based emotion benchmark for it. Also the proposed approach is applicable to the other works for emotion recognition. One may argue that conventional methods can work better for some specific cases. It is true that for some small dataset, the conventional methods works better than the deep learning methods. However, in the case where a large database is available, it is generally accepted that the deep learning could be more effective than conventional methods. Also, most of current research are focused on the employment of deep neural network. Therefore the presented method can be more easily combined with or utilized in the existing deep learning methods.

ACKNOWLEDGEMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058 & NRF-2017R1A4A1015559) and by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by Chonnam National University (Grant number: 2018-3293). The corresponding author is Guee-Sang Lee.

REFERENCES

[1] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey," IEEE Transactions on Affective Computing, vol. 4, no.1, 2013, pp. 15-33. doi: 10.1109/T-AFFC.2012.16

[2] M. Karg, A. Samadani, R. Gorbet, K. Kolja, J. Hoey, and D. Kulic, "Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation," IEEE Transactions on Affective Computing, vol. 4, no. 4, 2013, pp. 341-359. doi: 10.1109/T-AFFC.2013.29

[3] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. R. Scherer, "Toward a Minimal Representation of Affective Gestures," IEEE Transactions on Affective Computing, vol. 2, no. 2, 2011, pp. 106-118. doi: 10.1109/T-AFFC.2011.7

[4] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. R. Scherer, "Technique for Automatic Emotion Recognition by Body Gesture Analysis," Proc. CVPR, 2008. doi: 10.1109/CVPRW.2008.4563173

[5] S. Piana, A. Staglianò, F. Odone, A. Verri, and A. Camurri, "Real-time Automatic Emotion Recognition from Body Gestures," arXiv preprint arXiv:1402.5047, 2014.

[6] S. Piana, A. Staglianò, A. Odone, and A. Camurri, "Adaptive Body Gesture Representation for Automatic Emotion Recognition," ACM Transactions on Interactive Intelligent Systems, vol. 6, no. 1, 2016, p. 6. doi: 10.1145/2818740

[7] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," Proc. ICPR, 2006, pp. 1148-1153. doi: 10.1109/ICPR.2006.39

[8] H. Gunes and M. Piccardi, "Automatic Temporal Segment Detection and Affect Recognition from Face and Body Display," IEEE Transactions on Systems, Man, and Cybernetics, vol. 39, no. 1, 2009, pp. 64-84. doi: 10.1109/TSMCB.2008.927269

[9] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing Expression from Face and Body Gesture by Temporal Normalized Motion and Appearance Features," Jounal Image and Vision Computing. vol. 31, no. 2, 2011, pp. 175-185. doi: 10.1109/CVPRW.2011.5981880

[10] W. Wang and V. Enescu, "Adaptive Real-Time Emotion Recognition from Body Movements," ACM Transactions on Interactive Intelligent Systems, vol. 5, no. 4, 2016, p. 18. doi: 10.1145/2738221

[11] P. M. Muller, S. Amin, P. Verma, M. Andriluka, and A. Bulling, "Emotion Recognition from Embedded Bodily Expression and Speech during Dyadic Interactions," Proc. Affective Computing and Intelligent Interaction (ACII), 2016, pp. 663-669. doi: 10.1109/ACII.2015.7344640

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," Proc. of the IEEE, vol. 86, no. 11, 1998, pp. 2278-2324. doi: 10.1109/5.726791

[13] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," Neural Networks, 2015, pp. 140-151. doi: 10.1016/j.neunet.2015.09.009

[14] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” IEEE TPAMI, vol. 35, no. 1, 2013, pp. 221-231. doi: 10.1109/TPAMI.2012.59

[15] H. Ranganathan, S. Chakraborty, and S. Panchanathan, “Multimodal Emotion Recognition using Deep Learning Architectures,” Proc. IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-9. doi: 10.1109/WACV.2016.7477679

[16] S. T. Ly, G. S. Lee, S. H. Kim, and H. J. Yang, “3D Convolution and Convolution LSTM for Gesture-based Emotion Recognition,” Int. Symposium on Information Technology Convergence (ISITC), 2018.

[17] A. Torralba, R. Fergus, and W. Freeman, “80 Million Tiny Images: A Large Dataset for Non-Parametric Object and Scene Recognition,” IEEE TPAMI, vol. 30, no. 11, 2008, pp. 1957-1970. doi: 10.1109/TPAMI.2008.128

[18] R. Salakhutdinov and G. Hinton, “Semantic Hashing,” Int. J. of Approximate Reasoning, vol. 50, no. 7, 2009, pp. 969-978. doi: 10.1016/j.ijar.2008.11.006

[19] A. Andoni and P. Indyk, “Near-optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions,” IEEE Symposium on Foundations of Computer Science (FOCS'06), vol. 51, no. 1, 2008, pp. 117-122. doi: 10.1145/1327452.1327494

[20] Y. Weiss, A. Torralba, and R. Fergus, “Spectral Hashing”, Proc. NIPS, 2009, pp. 1753-1760.

[21] M. Raginsky and S. Lazebnik, “Locality Sensitive Binary Codes from Shift-Invariant Kernels,” Proc. NIPS, 2009, pp. 1509-1517.

[22] Y. Gong and S. Lazebnik, “Iterative quantization: A procrustean approach to learning binary codes,” IEEE TPAMI, vol. 35, no. 12, 2013, pp. 1916-2929. doi: 10.1109/TPAMI.2012.193

[23] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv preprint arXiv:1409.1556, 2014.

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” Proc. ICCV, 2015, pp. 4489-4497. doi: 10.1109/ICCV.2015.510

[25] R. Hou, C. Chen, and M. Shah, “Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos,” Proc. ICCV, 2017, pp. 5822-5831. doi: 10.1109/ICCV.2017.620

[26] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting,” International Conference on Neural Information Processing Systems, vol. 1, pp. 802-810.

[27] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, and T. Darrell, “Long-term Recurrent Convolutional Networks for Visual Recognition and Description,” IEEE TPAMI, vol. 39, no. 4, 2017, pp. 677-691. doi: 10.1109/TPAMI.2016.2599174

[28] German I. Parisi, Pablo Barros, Haiyan Wu, Guochun Yang, Zhenghan Li, Xun Liu, and Stefan Wermter, “A Deep Neural Model for Emotion-Driven Multimodal Attention,” The AAAI Spring Symposium on Interactive Multisensory Object Perception for Embodied Agents, 2017.

Son Thai Ly

He received the M.S in Material Engineering from Soongsil University, Korea in 2018. He currently is an M.S student at the Electronic and Computer Science department of Chonnam National University, Republic of Korea. His research interests are emotion recognition, deep learning, GAN, computer vision.

Guee-Sang Lee

He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Republic of Korea, in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering at Chonnam National University, Republic of Korea. His primary research interests are mainly in the field of image processing, computer vision, and video technology.

Soo-Hyung Kim

He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Republic of Korea, in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering at Chonnam National University, Republic of Korea. His primary research interests are mainly in the field of image processing, computer vision, and video technology.

Hyung-Jeong Yang

He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Republic of Korea, in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering at Chonnam National University, Republic of Korea. His primary research interests are mainly in the field of image processing, computer vision, and video technology.