SPATIOTEMPORAL ATTENTION BASED DEEP NEURAL NETWORKS FOR EMOTION RECOGNITION

Jiyoung Lee *Sunok Kim* *Seungryong Kim* *Kwanghoon Sohn*

School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: khsohn@yonsei.ac.kr

ABSTRACT

We propose a spatiotemporal attention based deep neural networks for dimensional emotion recognition in facial videos. To learn the spatiotemporal attention that selectively focuses on emotional sailient parts within facial videos, we formulate the spatiotemporal encoder-decoder network using Convolutional LSTM (ConvLSTM) modules, which can be learned implicitly without any pixel-level annotations. By leveraging the spatiotemporal attention, we also formulate the 3D convolutional neural networks (3D-CNNs) to robustly recognize the dimensional emotion in facial videos. The experimental results show that our method can achieve the state-of-the-art results in dimensional emotion recognition with the highest concordance correlation coefficient (CCC) on RECOLA and AV+EC 2017 dataset.

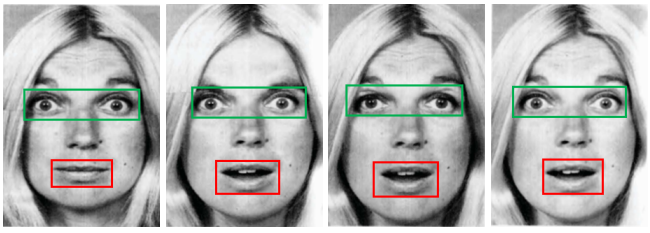
Index Terms— Emotion Recognition, Spatiotemporal attention, Convolutional Long Short-Term Memory, Recurrent Neural Network

1. INTRODUCTION

Emotion recognition has been one of the most important and fundamental problems in the development of interactive computer systems [1–3]. The ability to recognize facial expression and/or emotion is essential for a wide range of applications such as pain detection [4] and psychological distress [5].

Conventionally, most of efforts in emotion recognition [2, 6–9] have focused on *categorical* emotion description, where emotions are grouped into discrete categories such as surprise, fear, etc. [10, 11]. In the last few years, several methods have tried to recognize the six basic emotions using handcrafted [6–8] or learned [2, 9] feature based approaches. Although the state-of-the-art methods have shown satisfactory performance in categorical emotion recognition, those six basic emotions do not cover the full range of possible emotions, which hinders the application of emotion recognition methods to practical systems. An alternative way to represent emotions is *dimensional* emotion description [12], where emotions are described in a continuous domain, and *Arousal* and *Valence* are two representative domains. Arousal represents how engaged or apathetic a subject appears while valence represents how positive or negative a subject appears. Those models can represent more complex and subtle emotions with the higher-dimensional descriptions, which could be particularly useful for representing an untrimmed facial video data as exemplified in Fig. 1.

Over the past few years, deep convolutional neural networks (CNNs) based methods have shown substantially improved performance in emotion recognition tasks [1, 2, 9, 14]. However, most of those methods which use CNNs only cannot encode temporal information for a facial video sequence, and thus have shown limited performances for recognizing emotion in an untrimmed facial video. Although recurrent neural networks (RNNs) [1] and long short-term memory (LSTM) [15] have been used for understanding the facial video, they also have shown limited performances due to the lack of a mechanism for implicitly considering salient parts on the face.



(a) (b) (c) (d)

Fig. 1. Four type of surprise defined by Ekman [13] : (a) a questioning surprise, (b) astonished surprise, (c) dazed surprise, and (d) full surprise expression. Categorical emotion cannot cover full range of emotion (e.g., surprise represents the above four type of surprise), but dimensional emotion can represent subtle emotions. In addition, expression changes of surprise are represented in just two areas of the face, e.g., mouth (red box) and eye (green box), which exemplifies that estimating emotional sailent parts within facial videos is essential for recognizing emotion robustly.

On the other hands, facial action units (AUs) based emotion recognition approaches [9, 16] have found that only a small number of regions (*e.g.,* eyes, nose, and mouth) activate as a human changes their emotional expression [9], and attempt to detect AUs to estimate facial expression and recognize emotion by leveraging detected AUs [16]. However, those existing methods are based on the manual definition of individual AUs [9, 16], thus producing definite limitations in terms of providing optimal performance.

To overcome these issues, we propose a novel deep architecture that implicitly learn a spatiotemporal attention and estimate dimensional emotion for a facial video sequence. Specifically, we formulate a novel encoder-decoder network to learn spatiotemporal attention, where it first extracts the feature with spatial associations of each frame using 2D-CNNs and then estimates spatiotemporal attention using convolutional LSTM (ConvLSTM). Unlike conventional LSTM [17] is used to sequence learning [18], ConvLSTM enables us to maintain a spatial locality in the cell state while encoding the temporal correlation, and thus our attention inference module can estimate the attentive facial parts both spatially and temporally. Based on this spatiotemporal attention, the emotion recognition network is formulated using successive 3D convolutional neural networks (3D-CNNs) to deal with the sequential data. With a simple fusion scheme that convolutional activations from all frames within the input video are multiplied with estimated spatiotemporal attention, our network

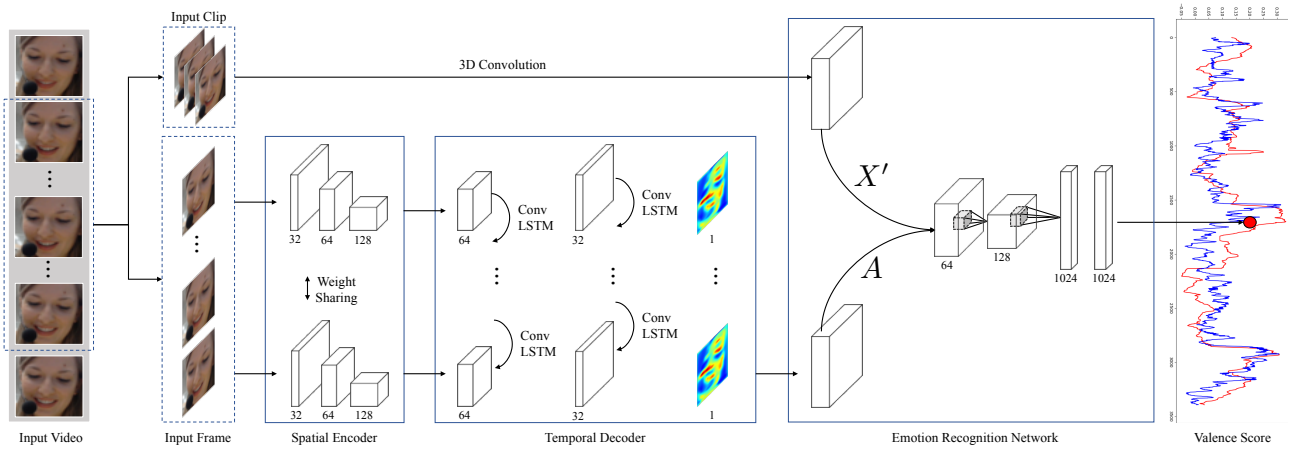


Fig. 2. The proposed learning framework for dimensional emotion recognition (valence score).

provides the state-of-the-art performance in the dimensional emotion recognition task for the facial video sequence.

2. PROPOSED METHOD

2.1. Problem Formulation and Overview

Let us define a facial video clip composed of a sequence of T frames as $I_{1:T} = \{I_1, I_2, ..., I_T\}$. The objective of dimensional emotion recognition is to regress a valence score $y \in [-1, 1]$ for each input frame $I_{1:T}$. We propose the novel learnable module that implicitly estimates *spatiotemporal attention* for the video. Our key-ingradient is to first extract the features of each frame with spatial associations using 2D-CNNs and then estimate spatiotemporal attention of the video using ConvLSTM (Section 2.2). The dimensional emotions of each frame are estimated by leveraging 3D-CNNs to encode both appearance and motion information simultaneously (Section 2.3). Fig. 2 shows the overall network configuration of our emotion recognition system.

2.2. Spatiotemporal Attention Network

We first introduce the attention inference network to predict a spatiotemporal attention for a facial video, which discovers emotional salient parts of the face. Since there is no supervision for the spatiotemporal attention, we design the inference network within a fully convolutional network in manner where the attention can be learned *implicitly* during learning the emotion recognition module only with the supervision of a valence label.

2.2.1. Spatial Encoder Network

Previous attention-based technique has learned attention by stack of LSTM (or RNNs) modules [18]. Although the method can employ temporal information, this model cannot take spatial correlation into consideration. To alleviate this limitation, we propose the feature encoder of 2D-CNNs. We extract convolutional feature activation $X_{1:T}$ for each frame $I_{1:T}$ within a Siamese network [19], where the weights and biases of each kernel are shared (*i.e.*, replicated across all frames and updated together during training phase), enabling us to reduce the number of parameters and prevent over-fitting problem. Specifically, the spatial encoder network consists of successive 3×3 convolution layers and rectified linear unit (ReLU) layers, followed by max-pooling layers with stride 2×2. To predict the attention with the same size of original images, those convolutional activations are enlarged through the temporal decoder network which will be described in Sec. 2.2.2.

2.2.2. Temporal Decoder Network

For convolutional features $X_{1:T}$ from the spatial encoder network, the temporal decoder network predicts the spatiotemporal attention for all T frames. The decoder network progressively enlarges the spatial resolution of $X_{1:T}$ through sequential deconvolutions similar to [19, 20]. Unlike other deconvolution layers as in [19, 20], we utilize ConvLSTM modules that encode the temporal correlation across inter-frames while preserving the spatial structure over sequences. Moreover, unlike LSTM that operates over sequences of vectors and performs biased linear transformations, ConvLSTM module has convolutional structures in both input-to-state and state-to-state transitions as follows:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} * c_{t-1} + b_i), \\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} * c_{t-1} + b_f), \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{sc} * X_t + W_{hc} * H_{t-1} + b_c), \\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot c_t + b_o), \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
\tag{1}
$$

where $\sigma(\cdot)$ and $\tanh(\cdot)$ are the logistic sigmoid and hyperbolic tangent (tanh) non-linearities, i_t, f_t, o_t, c_t and h_t are vectors to represent values of the input gate, forget gate, output gate, cell activation, and cell output at time t, respectively. $*$ denotes the convolution operator and $\odot$ denotes the Hadamard product. W_* are the filter matrices connecting different gates, and b_* are the corresponding bias vectors. The recurrent connections operate only over the temporal dimension, and use local convolutions to capture spatial context. With the ConvLSTM module, our temporal decoder network is composed of 3×3 ConvLSTM and tanh [21]. To enlarge the spatial resolution of X_t, we build the sequence of deconvolution with a factor of 2.

2.2.3. Spatiotemporal Attention Inference

Our spatiotemporal attention is used as a soft attention in a manner that this attention is multiplied to 3D convolutional feature activations. To this end, we first normalize the attention map spatially by

using the spatial softmax defined as follows [18]:

$$A_{t,i} = \frac{\exp(W_i^T H_{t-1})}{\sum_j \exp(W_j^T H_{t-1})} \quad i \in 1 \cdots H \times W, \tag{2}$$

where H_{t-1} is the hidden state, W_i are the weights mapping to the i^{th} element of the location softmax and j is defined for all locations. Through this spatial softmax, final spatiotemporal attention $A_{1:T}$ can be estimated. Note that our method does not need explicit predefined AUs and salient facial regions, and the attention inference module can be learned implicitly through the proposed network.

2.3. Emotion Recognition Network

By leveraging the spatiotemporal attention $A_{1:T}$, our method estimates a dimensional emotion for the facial video $I_{1:T}$. While the 2D-CNNs [1] can be used to predict the emotion for the facial video, it processes multiple input frames as different input channels independently, thus providing limited performances. To overcome this limitation, we employ the 3D-CNNs to deal with temporal information, which simultaneously consider spatial and temporal correlations across the input frames and directly regress the emotion.

To elegantly incorporate the spatiotemporal attention to emotion recognition through 3D-CNNs, we first extract convolutional feature activation $X'_{1:T}$ using 3D convolutional layers for the video $I_{1:T}$ as an input. Then, we multiply spatiotemporal attention $A_{1:T}$ to $X'_{1:T}$ to estimate the attention-boosted feature activations as follows:

$$X'' = A \odot X'. \tag{3}$$

For the attention-boosted feature activations X'', we finally formulate an additional 3D convolutional layers to infer dimensional emotion $\hat{y}$. This emotion prediction network has four 3D-convolution layers, three 3D max-pooling layers, and two fully-connected layers. The number of filters for four convolution layers are 32, 64, 128 and 256, respectively. The last fully-connected layer has a single output channels as f and we use a linear regression layer to estimate the output valence. We use the mean squared error as loss function. Our overall network can be learned only with a ground-truth valence label as a supervision.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Implementation Details

To recognize the emotion from a facial video, we first detected the face in each video frame using face and landmark detector in Dlib-ml [22], and then cropped the detected face region. We then mapped the detected landmark points to pre-defined pixel locations in order to normalize the eye and nose coordinates between adjacent frames.

We implemented our network using the TensorFlow library [23]. To reduce the effects of the network overfitting, we employed the dropout scheme with the ratio of 0.5 between fully-connected layer, and data augmentation schemes such as flips, contrast, and color changes. For training datasets, input videos in the training set were split into non-overlapped 16-frame clips. Thus, the input of model has a frame rate of 4 fps. For optimization, We chose Adam [24] due to its faster convergence than standard stochastic gradient descent with momentum. We trained our networks from scratch using mini-batches of 16 clips, with initial learning rate as $\lambda = 1e-4$. The filter weights of each layer were initialized by Xavier distribution, which was proposed by Glorot and Bengio [25], due to its properly scaled uniform distribution for initialization.

Table 1. Analysis on the performance of each component of the proposed network.

2D-CNN	3D-CNN	STA	RMSE	CC	CCC
✓			0.113	0.426	0.326
	✓		0.104	0.510	0.493
	✓	✓	0.102	0.572	0.546

Table 2. The qualitative evaluation of the predicted valence on RECOLA dataset [28]. The results with the lowest RMSE and highest CC/CCC were highlighted.

Method	RMSE	CC	CCC
Baseline [26]	0.117	0.358	0.273
CNN [1]	0.113	0.426	0.326
CNN + RNN ($\approx$ 1 sec.) [1]	0.111	0.501	0.474
CNN + RNN ($\approx$ 4 sec.) [1]	0.108	0.544	0.506
LGBP-TOP + LSTM [29]	0.114	0.430	0.354
LGBP-TOP + Bi-Dir. LSTM [15]	0.105	0.501	0.346
LGBP-TOP + LSTM + ϵ-loss [30]	0.121	0.488	0.463
CNN + LSTM + ϵ-loss [30]	0.116	0.561	0.538
3D-CNN + STA ($\approx$ 4 sec.)	**0.102**	**0.572**	**0.546**

Table 3. The qualitative evaluation of the predicted valence on AV+EC 17 dataset [31]. The results with the lowest RMSE and highest CC/CCC were highlighted.

Method	RMSE	CC	CCC
Baseline [31]	-	-	0.400
CNN [1]	0.114	0.564	0.528
CNN + RNN ($\approx$ 4 sec.) [1]	0.104	0.616	0.588
3D-CNN + STA ($\approx$ 4 sec.)	**0.099**	**0.638**	**0.612**

For all investigated methods, we interpolated the valence scores from adjacent frames related to dropped frames that the face detector missed. In addition, following the AV+EC's post-processing procedure of predictions [26, 27], we applied the same chain of post-processing on the obtained predictions; smoothing, centering and scaling except time-shifting.

3.2. Results

Experimental settings In order to evaluate the performance of the proposed method quantitatively, we computed three metrics: (i) Root Mean Square Error (RMSE), (ii) Pearson Correlation Coefficient (CC), and (iii) Concordance Correlation Coefficient (CCC) as used in [1]. Especially, the CCC tries to measure the agreement between two variables using the following expression:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4}$$

where ρ is the Pearson correlation coefficient, σ_x^2 and σ_y^2 are the variance of the predicted and ground truth values, and μ_x and μ_y are their means, respectively. The highest CC and CCC value represent the best recognition performance.

In the following, we evaluated our proposed network through comparisons to state-of-the-art CNNs-based approaches [1, 15, 29, 30]. The performance was measured on the RECOLA dataset [28], which has been adopted for the AudioVisual Emotion recognition Challenges (AV+EC) in 2015 [26] and 2016 [27]. We also evaluated our proposed method on the AV+EC 2017 dataset [31].

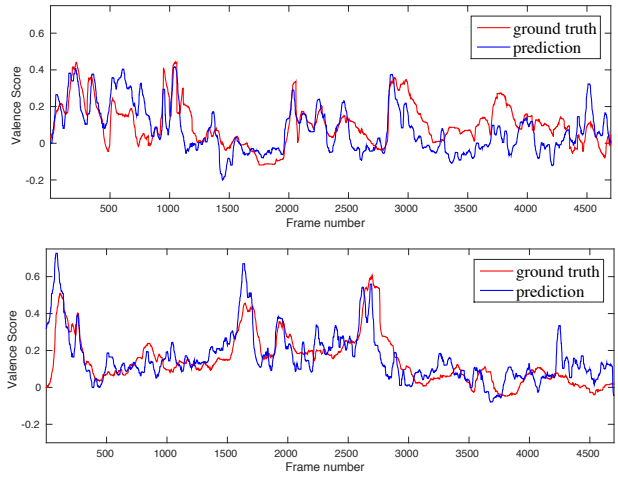


Fig. 3. Estimated valence graph of 5^{th} and 8^{th} subjects in development sets in RECOLA dataset [26].

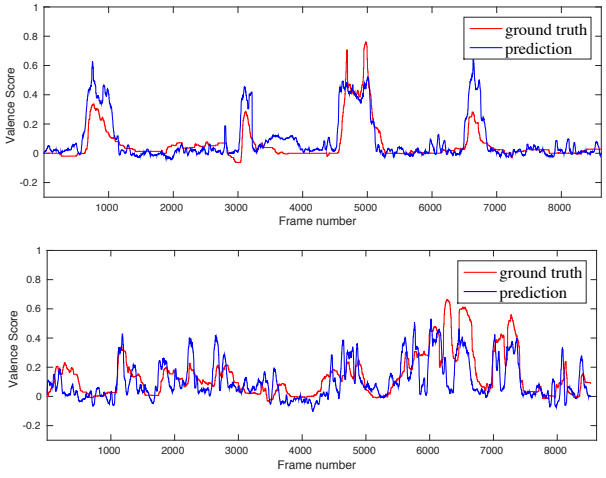


Fig. 4. Estimated valence graph of 4^{th} and 10^{th} subjects in development sets in AV+EC 2017 dataset [31].

Component-wise performance analysis We first evaluated the performance gain of each components in our method on the RECOLA dataset [28]. In order to analyze the effect of the proposed network architecture, we analyzed the performance of each component (i.e., encoder-decoder and 3D-CNNs) in Table 1. By learning the spatiotemporal attention using the encoder-decoder architecture, the estimation performance improves 0.062 and 0.053 for CC and CCC score compared than the performances using only 3D-CNNs which shows the effectiveness of proposed spatiotemporal attention based emotion recognition.

Visualization of attention maps To verify the effectiveness of the attention to estimate dimensional emotions, we visualized the attention maps where model focused on parts of the face, while improving the emotion recognition performance. As shown in Fig. 5, the proposed model effectively learn the important parts in the videos frames, especially eyes and mouth. At different frames, the proposed model captures different parts, since ConvLSTM deals with temporal correspondence. As a result, proposed attention network highlights relevant parts of emotion recognition and implicitly learn to detect specific AUs in facial images.

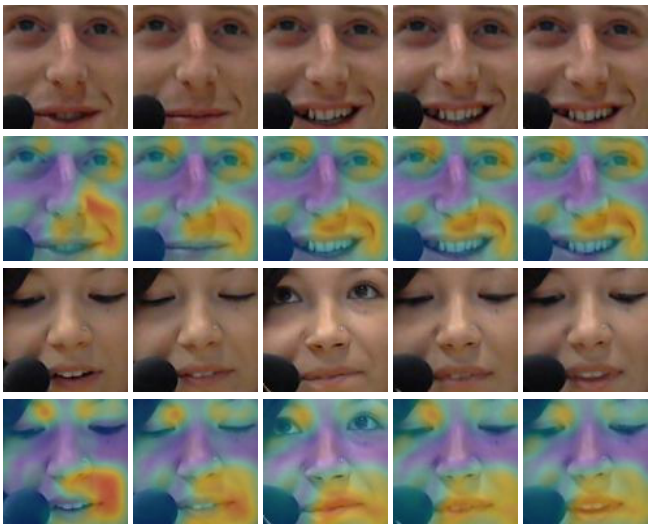


Fig. 5. Visualization of spatiotemporal attention maps learned by the proposed network in the RECOLA dataset [28]: Attention scores are normalized by the spatial softmax. Red indicates higher weight of the frame and blue indicates lower weight. Specifically, the areas around eyes and mouth are considered to be important to estimate emotion.

Comparison to other methods We then compared our method with the state-of-the-art methods including CNN-based approaches [1] and LSTM-based approaches [30] on the RECOLA dataset [28] in Table 2. The results showed that the proposed method exhibits a better recognition performance than conventional methods [1, 15, 29, 30].

In Table 3, we also compared our method with the RNN-based approach [1] on AV+EC 2017 dataset [31], which includes 34 training and 14 development videos. The results have also shown that the proposed method exhibits a better recognition performance compared to conventional methods.

We visualized the valence scores predicted by proposed method for three of the videos in the development set in Fig. 3 and Fig. 4. The proposed models can detect the valence score especially on the peak points by demonstrating the effectiveness of the proposed CNN architecture.

4. CONCLUSION

We proposed dimensional emotion recognition framework that leverages the spatiotemporal attention of video frames. Our method only considered spatial appearance and temporal motion for the facial video sequence simultaneously using 3D-CNNs. An extensive experimental analysis shows the benefits of our encoder-decoder attention network for dimensional emotion recognition, and state-of-the-art recognition performances on both RECOLA and AV+EC 2017 dataset. As future work, we will study a dimensional emotion recognition using multispectral database including RGB, depth, IR, and FIR to improve the emotion recognition performance.

5. ACKNOWLEDGMENTS

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7069370).

6. REFERENCES

[1] 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 *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 619–623.

[2] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, and R. C. Ferrari, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[3] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. IEEE Int. Conf. Face and Gesture Recognit.*, 2015, vol. 1, pp. 1–8.

[4] L. Patric, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. M . Prkachin, "Automatically detecting pain using facial actions," in *Proc. IEEE Affect. Comput. Intel. Inter. Work.*, 2009, pp. 1–8.

[5] J. M. Girard, J. F. Cohn, M. H. Mahoor, and S. M. Mavadati, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and Vis. Comput.*, vol. 32, no. 10, pp. 641–647, 2014.

[6] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 568–573.

[7] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *Proc. IEEE Int. Conf. Face and Gesture Recognit.*, 2006, pp. 223–230.

[8] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, 2007.

[9] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *Proc. IEEE Int. Conf. Comput. Vis. Work.*, 2015, pp. 19–27.

[10] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion.," *J. Personal. Social Psychology*, vol. 53, no. 4, pp. 712, 1987.

[11] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique.," 1994.

[12] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Research Personal.*, vol. 11, no. 3, pp. 273–294, 1977.

[13] P. Ekman and W. V. Friesen, "Unmasking the face: A guide to recognizing emotions from facial clues.," 2003.

[14] K. S. Ebrahimi, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 467–474.

[15] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. ACM Int. Work. Audio/Vis. Emot. Challenge*, 2015, pp. 73–80.

[16] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3304–3311.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] S. Sharma, R. Kiros, and r. Salakhutdinov, "Action recognition using visual attention," *arXiv:1511.04119*, 2015.

[19] J. Lee, H. Jung, Y. Kim, and K. Sohn, "Automatic 2d-to-3d conversion using multi-scale deep neural network," *in Proc. IEEE Int. Conf. Image Process.*, Sep. 2017.

[20] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn, "Depth prediction from a single image with conditional adversarial networks," *in Proc. IEEE Int. Conf. Image Process.*, Sep. 2017.

[21] S. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Neur. Inf. Proc. Syst.*, 2015, pp. 802–810.

[22] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[23] "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artific. Intell. Statis.*, 2010, pp. 249–256.

[26] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2015: The 5th international audio/visual emotion challenge and workshop," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1335–1336.

[27] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE Int. Conf. Face and Gesture Recognit.* IEEE, 2013, pp. 1–8.

[29] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. ACM Int. Work. Audio/Vis. Emot. Challenge*, 2015, pp. 49–56.

[30] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. ACM Int. Work. Audio/visual Emotion Challenge*, 2015, pp. 65–72.

[31] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi, and M. Pantic, "Avec 2017–real-life depression, and a ect recognition workshop and challenge," 2017.