Audio-Visual Attention Networks for Emotion Recognition

Jiyoung Lee
Yonsei University
Seoul, South Korea
easy00@yonsei.ac.kr

Sunok Kim
Yonsei University
Seoul, South Korea
kso428@yonsei.ac.kr

Seungryong Kim
Yonsei University
Seoul, South Korea
srkim89@yonsei.ac.kr

Kwanghoon Sohn*
Yonsei University
Seoul, South Korea
khsohn@yonsei.ac.kr

ABSTRACT

We present a spatiotemporal attention based multimodal deep neural networks for dimensional emotion recognition in multimodal audio-visual video sequence. To learn the temporal attention that discriminatively focuses on emotional sailient parts within speech audios, we formulate the temporal attention network using deep neural networks (DNNs). In addition, to learn the spatiotemporal attention that selectively focuses on emotional sailient parts within facial videos, the spatiotemporal encoder-decoder network is formulated using Convolutional LSTM (ConvLSTM) modules, and learned implicitly without any pixel-level annotations. By leveraging the spatiotemporal attention, the 3D convolutional neural networks (3D-CNNs) is also formulated to robustly recognize the dimensional emotion in facial videos. Furthermore, to exploit multimodal information, we fuse the audio and video features to emotion regression model. 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 AV+EC 2017 dataset.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*;

KEYWORDS

Multimodal emotion recognition, Spatiotemporal attention, Convolutional Long Short-Term Memory, Recurrent Neural Network

ACM Reference Format:
Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. 2018. Audio-Visual Attention Networks for Emotion Recognition. In *2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia (AVSU'18), October 26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3264869.3264873

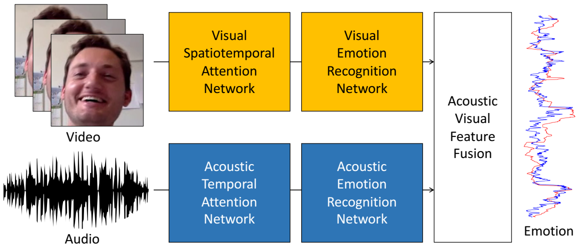

Video
Audio
Visual Spatiotemporal Attention Network
Visual Emotion Recognition Network
Acoustic Temporal Attention Network
Acoustic Emotion Recognition Network
Acoustic Visual Feature Fusion
Emotion

Figure 1: The proposed audio-visual attention and recognition networks for dimensional emotion recognition.

1 INTRODUCTION

Recognizing emotion in videos can facilitate a variety of interactive computer systems [13, 15, 27]. The ability to recognize facial expression and/or emotion is also essential for affective computing in artificial intelligence.

In order to represent emotion, dimensional emotion recognition is widely used, where *arousal* and *valence* are two representative domains described in a continuous domain. Specifically, *arousal* represents how engaged or apathetic a subject appears while *valence* represents how positive or negative a subject appears. Those models can represent subtle and complicated emotional behaviors.

To recognize the emotion using audio signals, a correlation between statistical measures of speech and the emotional state of the speaker is shown in previous works [2, 4]. Many acoustic features have been investigated for performing emotion classification, such as pitch-related features, energy-related features, Mel-frequency cepstrum coefficient, etc. In recent years, several audio features can be learned using deep neural networks (DNNs). In [16], dynamic time warping system is used to leverage the similarity to recognizie the affective label of the utterance. Rozgic et al. [23] performed emotion recognition by fusing acoustic features with lexical features extracted from DNN based emotion recognition system.

To extract useful features from the video sequence for emotion recognition, there exist many literatures. Over the past few years, deep convolutional neural networks (CNNs) based methods have shown substantially improved performances in emotion recognition tasks [5, 13–15]. However, most of those methods that only use CNNs 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

[*] Corresponding author

AVSU'18, October 26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5977-1/18/10.
https://doi.org/10.1145/3264869.3264873

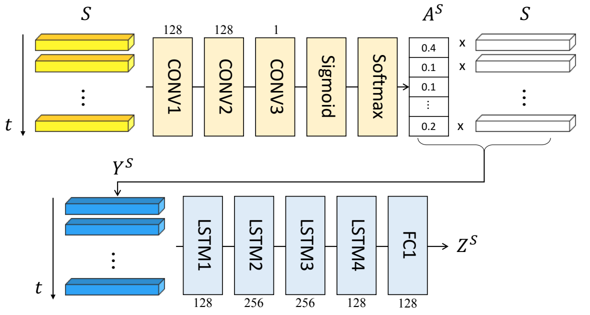


Figure 2: The proposed acoustic attention and recognition networks for dimensional emotion recognition.

neural networks (RNNs) [15] and long short-term memory (LSTM) [10] 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.

Recently, several methods have tried to recognize emotion by using not only visual features but also audio features, and have shown dramatically improved performance in emotion recognition [9, 25]. In this paper, we propose a novel deep architecture that implicitly learns a temporal attention for audio signals and a spatiotemporal attention for video signals, and estimates dimensional emotion (i.e., arousal and valence) in multimodal audio-visual video sequence. This paper extends the previous work of visual attention based emotion recognition system [20] by combining audio feature extraction and attention network. Specifically, we design a temporal attention network using DNNs to extract the most salient parts of audio signals. Also, we formulate a novel encoder-decoder network to learn the spatiotemporal attention in a manner that it first extracts the feature using 2D-CNNs and then estimates spatiotemporal attention using convolutional LSTM (ConvLSTM). Unlike conventional LSTM [11] is used to sequence learning [24], 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-CNNs to deal with the sequential data. To simultaneously use audio-visual information, the audio and video features are used as inputs of fusion network by concatenating the features. Our network provides the state-of-the-art performance in the dimensional emotion recognition task for the multimodal audio-visual database.

2 PROPOSED METHOD

The objective of our method is to recognize the dimensional emotion by simultaneously using audio-visual attentions and features. Let us define an audio sample composed of a sequence of T frames as $S_{1:T} = \{S_1, S_2, ..., S_T\}$ and a facial video sample 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 (or arousal) score $y \in [-1, 1]$ for each multimodal input frame $S_{1:T}$ and $I_{1:T}$. To accomplish this, we first extract audio features with its corresponding *temporal attention* (Section 2.1). Moreover, for video sequence, we propose the novel learnable module that implicitly estimates *spatiotemporal attention* for the video. We 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. The audio and video features are then fused using a late fusion method to benefit the complementary advantages of each modal (Section 2.3). Fig. 2 and Fig. 3 show the framework of acoustic and visual multimodal emotion recognition system.

2.1 Audio Temporal Attention Network

To extract acoustic features for emotion recognition, we introduce the temporal attention inference network which discovers emotional salient parts of the audio signals. By leveraging the baseline audio features in AV+EC 2017 database, we design the temporal inference network using DNNs, where the attention can be learned in a weakly-supervised manner, only with the supervision of a valence label. Fig. 2 shows the proposed audio temporal attention network.

2.1.1 Audio Feature Extraction Network. To extract audio features, AV+EC 2017 benchmark [22] adopts eGeMAPS as the baseline audio features. Concretely, both segment-level acoustic feature types are computed over segments of 4/6 seconds. Overall, the acoustic baseline feature has 88 dimensional features. The extraction of the LLDs and the computation of the funcionals are done using the openSMILE toolkit [6]. We use this feature for baseline acoustic feature.

2.1.2 Temporal Attention Inference. Many sequence learning [3, 29] propose an attention network to focus discriminative parts. We design the inference network within a deep neural networks (DNNs) without supervision for the temporal attention. The attention can be learned implicitly during learning the emotion recognition module which is consists of DNNs. Formally, let λ_t^S be the corresponding attention weight for S_t. We normalize the attention vector λ_t^S by using the temporal softmax as follows:

$$A_t^S = \frac{\exp(\lambda_t^S)}{\sum_t \exp(\lambda_t^S)} \quad t \in 1, \cdots, T. \tag{1}$$

Note that our method does not use ground-truth attention information to learn the acoustic attention inference module.

2.2 Visual Spatiotemporal Attention Network

To extract visual cues for emotion recognition, we introduce the attention inference network to predict 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 attention inference network within a fully convolutional network in a manner that the attention can be learned in a weakly-supervised manner, only with the supervision of a valence label. Fig. 3 shows the proposed visual spatiotemporal attention network.

2.2.1 Spatial Encoder Network. Previous attention-based approaches have learned attention by stack of LSTM (or RNNs) modules [24]. They only employ temporal information and does not

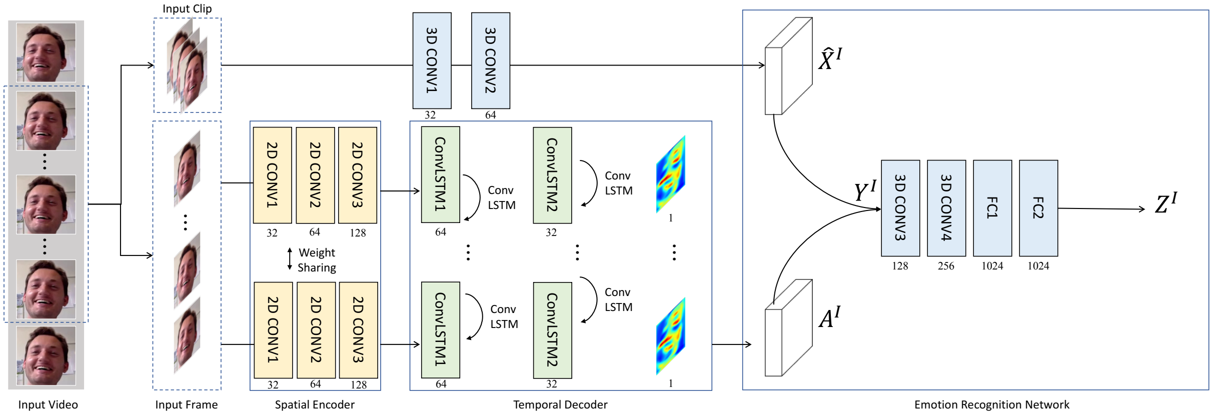


Figure 3: The proposed visual attention and recognition networks for dimensional emotion recognition.

consider spatial correlations. To alleviate this limitation, we propose the feature encoder of 2D-CNNs. We extract convolutional feature activation X_t for each frame I_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 an 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. From convolutional features X^I extracted in the spatial encoder network, the temporal decoder network learns the spatiotemporal attention for all T frames. The decoder network progressively enlarges the spatial resolution of X^I through a stack of deconvolution layers similar to [12, 19]. Unlike other deconvolution layers as in [12, 19], we use ConvLSTM modules that encode the temporal correlation across inter-frames while preserving the spatial structure. 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:

$$\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} * c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} * c_{t-1} + b_f),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{sc} * X_t + W_{hc} * H_{t-1} + b_c),\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot c_t + b_o),\\
h_t &= o_t \odot \tanh(c_t),
\end{aligned} \tag{2}$$

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 [28]. To enlarge the spatial resolution of X_t, we build the sequence of deconvolution with a factor of 2.

Table 1: Analysis on the performance of each component of the proposed network (in audio networks). 'LSTM-RNN' means acoustic features and 'TA' means acoustic temporal attention.

LSTM-RNN	TA	RMSE	CC	CCC
✓		0.122	0.496	0.454
✓	✓	0.109	0.621	0.551

Table 2: Analysis on the performance of each component of the proposed network (in video networks). 'STA' means visual spatiotemporal attention.

2D-CNN	3D-CNN	STA	RMSE	CC	CCC
✓			0.110	0.512	0.452
	✓		0.103	0.585	0.567
	✓	✓	0.099	0.638	0.612

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. Toward this end, we first normalize the attention map spatially by using the spatial softmax defined as follows [24]:

$$A^I_{t,i} = \frac{\exp(W_i^T H_{t-1})}{\sum_j \exp(W_j^T H_{t-1})} \quad i \in 1, \cdots, H \times W, \tag{3}$$

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^I can be estimated.

2.3 Multimodal Emotion Recognition Network

By leveraging the acoustic temporal attention A^S and visual spatiotemporal attention A^I, we recognize a dimensional emotion. We first use a soft attention mechanism to make attention-boosted

Table 3: The qualitative evaluation of the predicted valence on AV+EC 17 dataset [22]. The results with the lowest RMSE and highest CC/CCC were highlighted.

Audio	Video	Method	RMSE	CC	CCC
✓		Baseline[22]	-	-	0.351
	✓	Baseline[22]	-	-	0.400
	✓	CNN [15]	0.114	0.564	0.528
	✓	CNN + RNN (≈ 4 sec.) [15]	0.104	0.616	0.588
✓		LSTM-RNN + TA (≈ 4 sec.)	0.109	0.621	0.551
	✓	3D-CNN + STA (≈ 4 sec.)	0.099	0.638	0.612
✓	✓	**LSTM-RNN + TA + 3D-CNN + STA**	**0.091**	**0.664**	**0.641**

acoustic feature activations Y^S as follows:

$$Y^S = A^S \odot S \tag{4}$$

For the acoustic attention-boosted feature activation Y^S, we build additional 4 layers LSTM-RNN based neural network as speech emotion recognition network, which is a powerful tool for modeling sequential data [3, 10, 15]. In the last LSTM layer, we extract the acoustic feature for fusion with visual feature.

Moreover, for the facial video signals, 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 $\widehat{X}^I$ using 3D convolutional layers for the video I as an input. Then, we multiply spatiotemporal attention A^I to $\widehat{X}^I$ to estimate the attention-boosted feature activations as follows:

$$Y^I = A^I \odot \widehat{X}^I. \tag{5}$$

For the attention-boosted feature activations Y^I, we finally formulate an additional 3D convolutional layers to recognize dimensional emotion. 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. Before the last fully-connected layer has audio feature as Z^I, which and a linear regression layer is used to estimate the output valence.

To fuse the audio and video features, we modify the convolutional fusion scheme[7]. After the last fully connected layers of each emotion recognition network, we concatenate the acoustic features $Z^S \in \mathbb{R}^{d_S}$ and visual features $Z^I \in \mathbb{R}^{d_I}$, where d_S and d_I are dimensions of acoustic and visual features respectively, and subsequently convolve the stacked data Z with a fully connected layers where the output is a predicted emotion label for the last frame as follows:

$$Z = \text{cat}(Z^S, Z^I), \quad Z \in \mathbb{R}^{d_I + d_S} \tag{6}$$

To learn the networks, we use the mean squared error as loss function. It should be noted that our overall network can be learned only with a ground-truth valence label as a supervision.

3 EXPERIMENTAL RESULTS

3.1 Implementation Details

We implemented our network using the TensorFlow library [1]. To reduce the effects of the network overfitting, we employed the dropout scheme with the ratio of 0.5 between each fully-connected layer. For training datasets, input audio features in the training set were split into overlapped 16-frame feature map and input videos were also split into overlapped 16-frame clips. Thus, the input of model has a frame rate of 4 fps. For optimization, we chose Adam solver[18] 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 [8], due to its properly scaled uniform distribution for initialization.

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 [21, 26], we applied the same chain of post-processing on the obtained predictions; smoothing, centering and scaling except time-shifting.

3.2 Experimental Settings

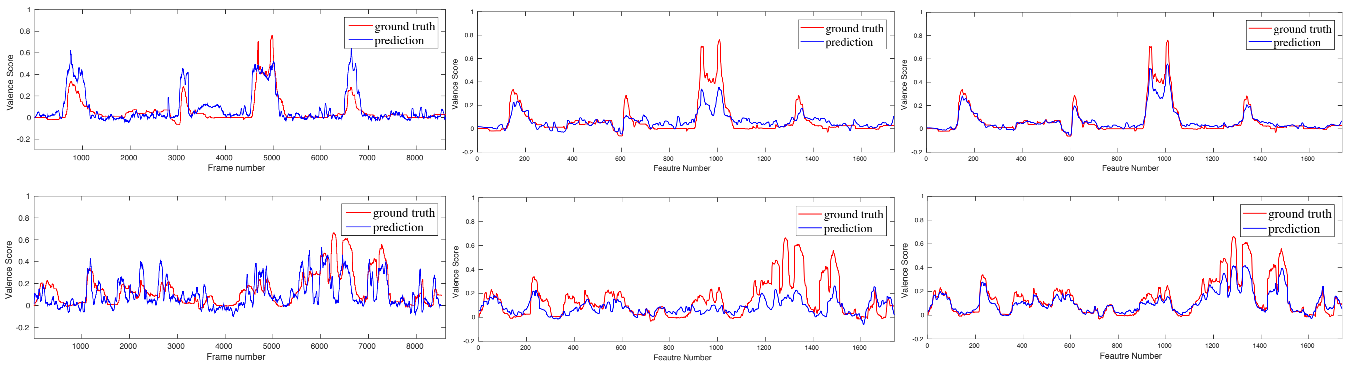
In order to evaluate the performance of the proposed method quantitatively, we computed three metrics: (i) Root Mean Square Error (RMSE), (ii) Pearson Correlation Coefficient (CC), and (iii) Concordance Correlation Coefficient (CCC) as used in [15]. The highest CC and CCC value represent the best recognition performance.

In the following, we evaluated our proposed network through comparisons to state-of-the-art CNNs-based approaches [15, 22]. The performance was measured on the AV+EC 2017 dataset [22], which has been adopted for the AudioVisual Emotion recognition Challenges (AV+EC) in 2017[22].

Likewise the other method [15], we use Dlib-ml [17] method as face and landmark detector. We then mapped the detected landmark points to pre-defined pixel locations in order to normalize the eye and nose coordinates between adjacent frames.

3.3 Results

3.3.1 Component-wise Performance Analysis. We evaluated the performance gain of each components in our method on the AV+EC



(a) (b) (c)

Figure 4: Estimated valence graphs of 4^{th} and 10^{th} subjects in development sets in AV+EC 2017 dataset [22]: (a) estimated graph on video feature. (b) estimated graph on audio feature. (c) estimated graph on both audio and video features.

2017 dataset [22]. In order to analyze the effect of the proposed network architecture, we analyzed the performance of each component (i.e., LSTM-RNN, attention DNN, encoder-decoder and 3D-CNNs) in Table 1 and Table 2. Acoustic temporal attention improves performance 0.125 and 0.097 for CC and CCC score compared than the performance using only LSTM-RNN architecture. On the other hands, by learning the spatiotemporal attention using the encoder-decoder architecture, the estimation performance improves 0.053 and 0.045 for CC and CCC score compared than the performances using only 3D-CNNs which shows the effectiveness of proposed spatiotemporal attention based emotion recognition.

3.3.2 Comparison to Other Methods. In Table 3, we then compared our method with the RNN-based approach [15] on AV+EC 2017 dataset [15, 22], 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 [15, 22].

We compared the valence scores predicted by proposed method to ground-truth valence labels for two of the videos in the development set in Fig. 4. The proposed models can detect the valence score especially on the peak points by demonstrating the effectiveness of the proposed attention architecture.

4 CONCLUSIONS

We proposed the dimensional emotion recognition framework that leverages both acoustic temporal attention and visual spatiotemporal attention of multimodal videos. Our method considered spatial appearance and temporal motion for the facial video sequence simultaneously using 3D-CNNs, while attention DNNs are implicitly focused on temporal acoustic saliency parts. Finally, we fused the features extracted from audio-visual domains. An extensive experimental analysis shows the benefits of our attention network for dimensional emotion recognition, and demonstrates state-of-the-art recognition performances of our method on the AV+EC 2017 dataset.

5 ACKNOWLEDGEMENTS

This research was supported by the International Research Development Program of the National Research Foundation of Korea (NRF) funded by Ministry of Science and ICT (NRF-2017K1A3A1A16066838).

REFERENCES

[1] 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/ Software available from tensorflow.org.

[2] N. Amir and S. Ron. 1998. Towards an automatic classification of emotion in speech. In *Proc. Conf. Int. Speech Communicat. Associat.*

[3] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*. 577–585.

[4] F. Dellaert, T. Polzin, and A. Waibel. 1996. Recognizing emotion in speech. In *Proc. Conf. Int. Speech Communicat. Associat.*

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

[6] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[7] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1933–1941.

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

[9] H. Gunes, M. Pantic, and A. S. Ashour. 2010. Automatic, Dimensional and Continuous Emotion Recognition. In *Int. Journal of Synthe. Emotio.*, Vol. 1. 68–99.

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

[11] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn. 2017. Depth Prediction from a Single Image with Conditional Adversarial Networks. *in Proc. IEEE Int. Conf. Image Process.* (Sep. 2017).

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

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

[15] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang. 2016. How deep neural networks can improve emotion recognition on video data. In *Proc. IEEE Int. Conf. Image Process.* 619–623.

[16] Y. Kim and E. M. Provost. 2013. Emotion Classification via Utterance-level Dinamics: A Pattern-based appraoch to characterizing aaffective expressions. In *IEEE Int. Conf. Acous, Speech and Signal Process.* 3677–3681.

[17] D. E. King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[18] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).

[19] J. Lee, H. Jung, Y. Kim, and K. Sohn. 2017. Automatic 2D-to-3D Conversion using Multi-scale Deep Neural Network. *in Proc. IEEE Int. Conf. Image Process.* (Sep. 2017).

[20] J. Lee, S. Kim, S. Kim, and K. Sohn. 2018. Spatiotemporal Attention Based Deep Neural Networks for Emotion Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

[22] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi, and M. Pantic. 2017. AVEC 2017–Real-life Depression, and A ect Recognition Workshop and Challenge. (2017).

[23] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad. 2012. Emotion recognition using acoustic and lexical features. In *Proc. Conf. Int. Speech Communicat. Associat.*

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

[25] B. Sun, S. Cao, and L. Li. 2015. Exploring multimodal visual features for continuous affect recognition. In *Proc. ACM Int. Work. Audio/Vis. Emot. Challenge*. 83–88.

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

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

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

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.