Deep, dimensional and multimodal emotion recognition using attention mechanisms

Jan Lucas, Esam Ghaleb, and Stylianos Asteriadis[0000−0002−4298−6870]

Department of Data Science & Knowledge Engineering, Maastricht University,
Paul-Henri Spaaklaan 1, 6229 EN Maastricht, Netherlands
jan-lucas@hetnet.nl
{esam.ghaleb, stelios.asteriadis}@maastrichtuniversity.com

Abstract. Emotion recognition is an increasingly important sub-field in artificial intelligence (AI). Advances in this field could drastically change the way people interact with computers and allow for automation of tasks that currently require a lot of manual work. For example, registering the emotion a subject expresses for a potential advert. Previous work has shown that using multiple modalities, although challenging, is very beneficial. Affective cues in audio and video may not occur simultaneously, and the modalities do not always contribute equally to emotion. This work seeks to apply attention mechanisms to aid in the fusion of audio and video, for the purpose of emotion recognition using state-of-the-art techniques from artificial intelligence and, more specifically, deep neural networks. To achieve this, two forms of attention are used. Embedding attention applies attention on the input of a modality-specific model, allowing recurrent networks to consider multiple input time steps. Bimodal attention fusion applies attention to fuse the output of modality-specific networks. Combining both these attention mechanisms yielded CCCs of 0.62 and 0.72 for arousal and valence respectively on the RECOLA dataset used in AVEC 2016. These results are competitive with the state-of-the-art, underlying the potential of attention mechanisms in multimodal fusion for behavioral signals.

Keywords: Emotion Recognition · Multimodal · Neural Networks · Attention Mechanisms

1 Introduction

Emotion recognition as a field in machine learning tries to automate the identification of emotion in a human subject through various means, including computer vision, signal processing and deep learning. This problem is non-trivial as emotion in itself is an abstract concept that is hard to interpret, and thus many different models have been proposed to describe it [9]. Interpretation of emotional expressivity is hindered by the differences in expression between cultures and even persons. Emotion recognition related experiments and data, often come in the form of two types: acted and spontaneous. In the former category, emotions are many times expressed by an actor, while the spontaneous category mostly involves video clips of spontaneously expressed emotions. Spontaneous emotion recognition is deemed to be harder since it deals with more genuine expression of emotion, which tends to be more subtle than in the acted case.

The state-of-the-art approaches for emotion recognition make use of multiple modalities. This entails that emotion will be predicted by looking at multiple sources. Emotion can be expressed by, for example, both facial and vocal expression, and taking both of these sources into account leads to better performing models [7, 11, 4]. However, using multiple modalities is challenging. These modalities usually differ in multiple aspects, such as their inherent distributions, synchronization, sampling rate, dimensionality, etc.

This work aims to make use of the effectiveness of transfer learning by utilizing pre-trained networks on multiple modalities and fusing these using attention mechanisms. For this purpose, a new method is proposed, that combines previous work in emotion recognition and neural attention to predict emotions in the valence-arousal emotion spectrum.

1.1 Related Work

Work by Ghaleb et al. [4] proposed a framework, which uses metric learning and combines multiple input modalities that showed an increase in performance for discrete emotion recognition in an acted setting [4]. This work relies on some of the preprocessing techniques developed and tested by Ghaleb et al., since it is expected that they will also be beneficial for continuous emotion recognition. Research in the field of emotion recognition is encouraged by the Audio Visual Emotion Challenge (AVEC), which is a competition that is held yearly as part of the ACM Multimedia conference. The competition features a continuous affect recognition sub-part that sets a benchmark for work in the emotion recognition field. The accompanying baseline uses support vector machines (SVM) for each of the eight used modalities, with modality specific feature extraction and processing. SVMs are subsequently fused with a linear regression model [10]. This work will focus on solving the problem presented in this sub-part using feature embeddings and deep neural networks, which have been shown to be effective in similar works [10, 12, 3, 6]. Furthermore, Wu et al. showed that using only feed-forward networks and attention can achieve results that are comparable to recurrent approaches [11]. The method proposed in this work relies on the architectures used by Zhao et al.[12] and Haifeng et al.[3], in which they show that using stacked LSTMs followed by a dense linear layer per modality provides good results. Brady et al. showed state-of-the-art results on the RECOLA dataset for AVEC 2016 using a Kalman Filter approach to fuse models trained for specific modalities [2].

2 Methods

2.1 Attention

Attention mechanisms consitute a family of techniques that can be applied to selectively focus on parts of a sequence. It was initially proposed by Luong et al. [8] for the purpose of machine translation using an encoder-decoder network. Here the encoder network encodes the sentence in the source language and the decoder uses this encoding to reproduce the sentence in the target language. The order of the words in the source and target languages are often not aligned, and thus the decoder needs to be able to process the encoding out of order. Attention mechanisms allow this network to selectively focus on relevant parts of the encoding to produce a part of the target sequence. The attention mechanism, specifically the general-dot-product (GDP) variant, calculates an attention vector a_t over each encoder hidden state $\overline{h}_s$, which determines how important each part of the input is. This mechanism is formulated as follows:

$$a_t(s) = \frac{\exp(\text{score}(h_t, \overline{h}_s)}{\sum_{s'} \exp(\text{score}(h_t, \overline{h}_{s'})} \tag{1}$$

$$\text{score}(h_t, \overline{h}_s) = h_t^\top W_a \overline{h}_s \tag{2}$$

Here $h_t^\top$ is the current state of the decoder and matrix W_a is a parameter that is to be learned. The attention vector a_t can then be used to compute a weighted average of the source hidden states [8]. Similar to translation, emotions and their corresponding cues in the data may not be aligned. Thus the use of attention mechanisms could improve the performance of models for emotion recognition by allowing them to focus on the information that is relevant for recognizing emotion.

2.2 Proposed approach

The proposed model follows the works of Zhao et al. and Haifeng et al. [12, 3]. These works successfully use LSTMs for multi-modal affect recognition using features that are similar to the ones used in this work. The model used here is a stacked LSTM with dropout between the layers combined with a dense linear layer, where each model outputs both valence and arousal for each time step. This architecture is called the deep long short-term memory network (DLSTM). In total, the network consists of two LSTM layers followed by a dense linear layer. The first LSTM layer iterates over the input sequence and produces a hidden state for each time step in the input. These hidden states are the input for the second LSTM layer, which in turn produces a new sequence of hidden states. These final hidden states are the input to the linear layer that maps them into the two emotion dimensions. This architecture is expanded with attention-based layers on the embedding level (over time) and the fusion level (over the DLSTM networks). Figure 1 shows the overall network structure.

Embedding attention Attention over the embeddings closely follows the general dot-product (GDP) approach described by Luong et al. for language translation using encoder-decoder architectures [8]. The GDP attention mechanism computes scores for each source hidden state $\overline{h}_s$ in a sequence. A softmax function is applied to these to obtain a distribution which is subsequently used to construct a weighted combination of the sequence. GPD attention computes the score in the following way: $h_t^\top W_a \overline{h}_s$. Where h_t is the current state of the decoder and $\overline{h}_s$ is a hidden state of the encoder. W_a is a weight matrix that maps the encoder hidden state and the target hidden state to a score and has to be optimized. To adapt this method for affect recognition, the target hidden state h_t is the hidden state of the DLSTM, C_v in figure 1, and the source states are replaced with the embedding vectors in a certain window of size n, represented by $V_{t-\frac{n}{2}} \cdots V_{t+\frac{n}{2}}$ in figure 1. The attention mechanism thus computes the score of an embedding vector depending on the current state of the DLSTM and the contents of the embeddings.

This attention mechanism is applied only to the video modality. As will be explained in section 3, the extracted audio features already contain temporal information and should therefore benefit less from attention.

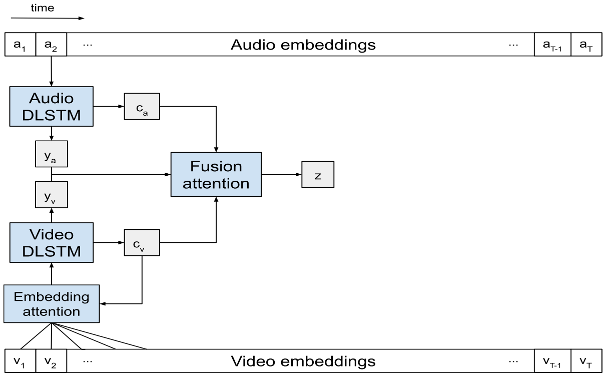

Fig. 1. Overview of the proposed model. a_t and v_t respectively, are the audio and video embeddings at time t. The variables y_m and c_m are the output and hidden state of the DLSTM responsible for modality m, and z represents the fused output.

Bimodal attention fusion The attention concept is also applied to decision fusion. The information that the DLSTMs receive may not allow for a very good prediction of emotion at every time step. For example, at some time t, audio may be more appropriate for emotion recognition, whereas the visual signal may not be carrying significant information. In that case, the output of the audio LSTM should have a higher weight in decision making. GDP attention mechanism outputs a linear combination of the inputs, resulting in equal weight being given to the valence and arousal output of a DLSTM. This is not necessarily the best solution, as at some time step t the audio DLSTM might be very "confident" about its output for arousal and unsure for valence, but the video DLSTM may be confident about valence. The GDP attention mechanism cannot assign different weights for each output dimension separately.

We experimented with an attention approach to decision fusion: Bimodal Attention Fusion. This method uses attention to combine the outputs of modality-specific DLSTM networks. Attention scores are computed by considering all DLSTM hidden states at once. Here $C \in \mathbb{R}^{cn}$, with c being the size of the hidden states and n the number of DLSTMs, is a vector containing the concatenation of the hidden states of all the underlying DLSTMs at time t. $Y_m \in \mathbb{R}^n$ is a vector containing the outputs of the DLSTMs for output modality m. This vector contains either valence or arousal outputs of all the DLSTM networks. The scores and outputs are computed separately for both arousal and valence, because different output dimensions may require different distribution of attention. The calculation of the scores can be reformulated as follows:

$$S_m = C^\top W_m \tag{3}$$

$$a_{m_i} = \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)} \tag{4}$$

$$Z = \{z_1, \cdots, z_o\}, \text{where } z_m = \sum_{i=1}^{n} a_{m_i} y_{m_i} \tag{5}$$

The attention vector A, whose elements are described in equation 4, is the matrix product of the DLSTM hidden states in C and the learned weights in W_m followed by an application of the softmax function. The weights $W_m \in \mathbb{R}^{cn \times n}$ map the hidden states to scores per DLSTM and are optimized separately per output modality m. This allows for different mappings from hidden states to scores for both valence and arousal. The attention weights in A_m are used to take a weighted combination of the DLSTM outputs.

Baseline fusion methods The effectiveness of the attention mechanism is assessed by comparing it to networks that fuse without attention. The first network realizes fusion as a static combination of the network outputs by a dense linear layer and is named output linear baseline (OLB). When $Y \in \mathbb{R}^{o \times n}$ is a matrix containing the LSTM outputs, with o being the number of output dimensions, and $W \in \mathbb{R}^{n \times 1}$ is a weight matrix, the fused output can be formulated as follows:

$$Z = YW \tag{6}$$

Matrix W is optimized during training and linearly fuses the outputs of the DLSTMs. The second network fuses by using the concatenation of the hidden states of the DLSTMs. This method, named hidden linear baseline (HLB), can be formulated as:

$$Z = C^\top W \tag{7}$$

Just as in equation 3, $C \in \mathbb{R}^{cn}$ is a matrix containing the concatenation of the hidden states of the DLSTMs, but here $W_m \in \mathbb{R}^{cn \times o}$ is a parameter that maps the hidden states directly to the output emotions.

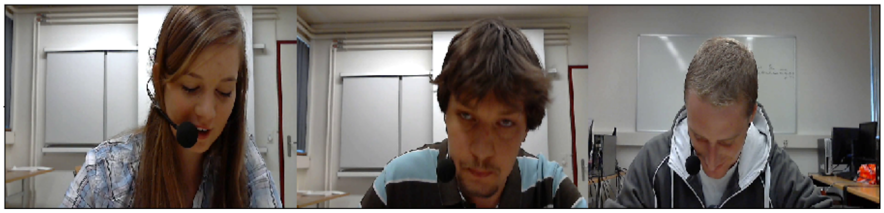

Fig. 2. Example frames from the RECOLA dataset

3 Results and discussion

This section details the experiments performed to determine the performance of the architecture described in section 2.2. A subset of the RECOLA dataset from the University of Fribourg was used in the 2015 and 2016 Audio/Visual Emotion Challenge (AVEC), and is used in this work to train the proposed architecture. The dataset consists of 18 five minute clips in predetermined train and validation partitions [10]. Examples of frames from this dataset are shown in figure 2. This subset of RECOLA contains several feature types, but only the raw video and audio data is used. Emotion is annotated continuously in the valence-arousal space for each video frame. Face extraction is performed on each frame of video and the raw data is transformed using embedding networks. Activations from the last convolutional layer of the VGGFace network is used to extract frame-level video features and VGGish is used to extract audio features from a 960ms window [1, 5].

The models mentioned below are trained by optimizing the mean squared error using the Adam optimizer with a learning rate of 0.01. The DLSTM architecture, described in section 2.2, is optimized using truncated backpropagation through time, following Zhao et al. and Haifeng et al. [12, 3]. The training and test partitions provided with the RECOLA data were used in order to make a fair comparison with the results from the state of the art from AVEC. Model performance is assessed with the Concordance Correlation Coefficient (CCC) measure, which is a common measure of performance in emotion recognition. The CCC is computed for each sequence in test partition of the dataset and averaged over the sequences to show the performance.

3.1 Attention

As explained in section 2.2, attention mechanisms are applied on two points in the proposed model. Embedding attention, which applies attention to the embeddings used as input to the video DLSTM, and fusion attention, which uses attention to fuse the outputs of the DLSTMs. These two methods are evaluated separately in the following experiments.

Embedding attention To assess the effectiveness of applying attention to the video embeddings, the performance of the DLSTM is compared with and without attention. For embedding attention, a window size of 13 frames is used, allowing the mechanism to consider a segment of half a second. The baseline in this comparison is a DLSTM without any attention, and thus processes the embeddings sequentially, instead of being able to focus on a window.

Initial performance of the model with embedding attention was very poor and stagnated directly after start of training. This was in contrast with the behavior of the model without embedding attention and suggested that training was hindered by the addition of the attention mechanism. To counteract this, the embedding attention layer is bypassed for the first five training epochs. After this startup period, the embedding layer is included again, which significantly improved performance. A possible explanation for this phenomenon is the cyclic dependency between the embedding attention and the hidden layer of the DLSTM. The embedding attention layer uses the hidden state to compute the input to the DLSTM, which in turn affects the hidden state. A meaningless hidden layer could result in poorly attended input, which then maintains the form of the hidden state. To account for random initialization, training and testing is repeated 10 times. For arousal, this resulted in very similar results regardless of the use of attention, with CCCs of 0.15 (±0.05) and 0.14 (±0.05) for no attention and embedding attention respectively. A slight improvement was observed for valence with CCC results of 0.36 (±0.07) without attention and 0.39 (±0.06) with embedding attention. Even though promising, this difference is not statistically significant, with $p > 0.05$. A bidirectional variant of the DLSTM without attention, since the embedding attention allows the network to use frames ahead of the current time step, and a bigger attention window were evaluated, but these resulted in similar CCC scores.

Table 1. CCC results for the pre-trained uni-modal DLSTMs (left) and their fusion using bimodal attention fusion and baselines (right). Both Bimodal attention and Hidden linear baseline (HLB) successfully fuse the unimodal networks for valence prediction.

Uni-modal	Valence	Arousal
audio	0.42	0.60
video	0.24	0.10

Fusion	Valence	Arousal
Bimodal attention	0.48 (±0.04)	0.60 (±0.1)
OLB	0.32 (±0.03)	0.40 (±0.08)
HLB	0.48 (±0.01)	0.64 (±0.02)

Fusion attention Section 2.2 describes the attention mechanism that can be used to combine the outputs of the uni-modal DLSTM networks. To make a fair comparison with the detailed baselines, two DLSTM networks are pre-trained separately on audio and video, and subsequently frozen before training the fusion mechanisms. This procedure restricts the performance of the model as a whole, but allows for a clear comparison of the fusion methods. The training of the fusion mechanisms is repeated 10 times while using the same pre-trained DLSTMs, to account for random initialization.

The results can be seen in figure 3 and table 1. The CCC values for the pre-trained network are also detailed in table 1. Comparing the methods shows that the proposed attention fusion mechanism significantly improves performance when compared to fusion by linearly combining the DLSTM outputs (OLB). However, its performance is matched by the baseline method that regresses the hidden states of the DLSTMs directly (HLB). For valence, the HLB baseline and bimodal attention mechanism both showed better performance than the unimodal networks they fused. This suggests that the performance of the audio network is slightly increased by fusion with the video modality, but the difference is not large enough for any concrete conclusions. In short, the proposed method seems fuse the uni-modal networks successfully, however its performance does not improve beyond the HLB baseline, which does not use attention.

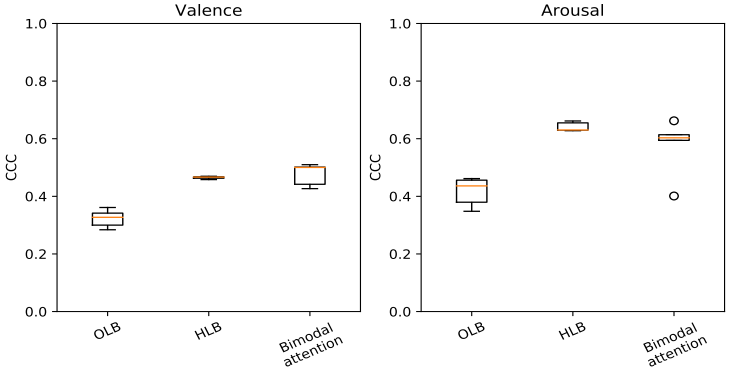

Fig. 3. CCC results for fusion of pre-trained uni-modal DLSTM models using bimodal attention fusion and baseline methods.

3.2 Comparison with the state of the art

The previous sections have each evaluated parts of the proposed architecture. In this section, a comparison with recent works from the literature is performed. For this, the network is trained in an end-to-end fashion using the bimodal attention fusion method and embedding attention on the video modalty. Hyperparameters are optimized empirically, resulting in hidden sizes of 32 and 128 for the audio and video DLSTMs respectively. Work by Haifeng et al. [3] shows that early fusion of features combined with decision level fusion provides improved results for emotion recognition. For this purpose, audio and video features are concatenated per time step to form an early fusion modality. The model described in section 2.2 is extended with a third DLSTM, with a hidden size of 128, that is used on this new modality. The outputs of this DLSTM are fused with the outputs from the audio and video DLSTMs to produce the final model output. The results are compared with a baseline provided by AVEC [10] and the best results on this dataset, achieved by Brady et al. [2]. This comparison is displayed in table 2. The CCC scores obtained by the proposed model for arousal are higher than the baseline, but are below the ones by Brady et al. However, it should be highlighted here that these two works make use of a wider set of modalities (such as electrodiograms and electrodermal activity, beyond just video and audio), whereas, in the proposed method, only audio and video are considered. For valence, the performance is just below the baseline. The results obtained with the proposed model are comparable to the results of these methods, even though fewer modalities were used and embedding techniques from other domains were reused.

Table 2. Performance of the model proposed in section 2.2 compared to the AVEC baseline and state of the art for this dataset.

	Valence	Arousal
Baseline	0.683	0.639
State of the art (Brady et al.)	0.702	0.82
Proposed	0.62	0.72

4 Conclusions and future work

This work explored combining the information in audio and video data by using attention to fuse the output of networks that were trained on only one modality each. Furthermore, attention was used to spot important video embeddings in temporal windows using the hidden state of an LSTM network.

Fusion of the output modalities using attention shows a significant improvement when compared to a model that does not take the states of the input networks into account. However, it shows similar performance to the baseline that directly regresses the hidden states, suggesting that more improvements should be possible. Usage of embedding attention showed promising results, but this difference is not significant, with a p-value greater than 0.05. Applying attention on the embedding level produced new challenges, that were overcome by using a special training procedure. Future work could investigate the causes for this and explore other, more flexible, variants of this mechanism.

Combining embedding attention and fusion attention yields a model that shows promising performance. Results exhibited improved performance compared to the AVEC baseline for arousal and close performance for valence. The proximity to the baseline and state-of-the art results shows the potential of the proposed method, since the baseline and state-of-the-art methods use more modalities and fine-tuned pre-processing methods. This is in contrast to the proposed method, which uses fewer modalities and reuses feature embeddings from other domains. Other modalities can be easily included in the proposed method and it is expected that this will improve results.

In conclusion, the use of attention mechanisms for emotion recognition shows promising results and can successfully combine information from multiple modalities. Future research could expand on this architecture by experimenting with different forms of attention, extra modalities and different feature embeddings.

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. p. 279–283 (2016). https://doi.org/10.1145/2993148.2993165
2. Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., Huang, T.S.: Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. p. 97–104. AVEC '16, Association for Computing Machinery (2016). https://doi.org/10.1145/2988257.2988264
3. Chen, H., Deng, Y., Cheng, S., Wang, Y., Jiang, D., Sahli, H.: Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. p. 19–26. AVEC '19, Association for Computing Machinery (2019). https://doi.org/10.1145/3347320.3357690
4. Ghaleb, E., Popa, M., Asteriadis, S.: Metric learning based multimodal audio-visual emotion recognition. IEEE MultiMedia pp. 1–1 (2019). https://doi.org/10.1109/MMUL.2019.2960219
5. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 131–135
6. Huang, J., Tao, J., Liu, B., Lian, Z., Niu, M.: Efficient modeling of long temporal contexts for continuous emotion recognition. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 185–191 (9 2019). https://doi.org/10.1109/ACII.2019.8925452
7. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B.W., Star, K., Hajiyev, E., Pantic, M.: SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2019)
8. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421. Association for Computational Linguistics (Sep 2015). https://doi.org/10.18653/v1/D15-1166
9. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37** (02 2017). https://doi.org/10.1016/j.inffus.2017.02.003
10. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. p. 3–10. AVEC '16, Association for Computing Machinery (2016). https://doi.org/10.1145/2988257.2988258
11. Wu, Z., Zhang, X., Zhi-Xuan, T., Zaki, J., Ong, D.C.: Attending to emotional narratives. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 648–654. IEEE Computer Society (sep 2019). https://doi.org/10.1109/ACII.2019.8925497
12. Zhao, J., Li, R., Chen, S., Jin, Q.: Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. p. 65–72. AVEC'18, Association for Computing Machinery (2018). https://doi.org/10.1145/3266302.3266313