MULTIMODAL ATTENTION-MECHANISM FOR TEMPORAL EMOTION RECOGNITION

Esam Ghaleb, Jan Niehues, and Stylianos Asteriadis

Maastricht University, Maastricht, the Netherlands
{esam.ghaleb, jan.niehues, stelios.asteriadis}@maastrichtuniversity.nl

ABSTRACT

Exploiting the multimodal and temporal interaction between audio-visual channels is essential for automatic audio-video emotion recognition (AVER). Modalities' strength in emotions and time-window of a video-clip could be further utilized through a weighting scheme such as attention mechanism to capture their complementary information. The attention mechanism is a powerful approach for sequence modeling, which can be employed to fuse audio-video cues over-time. We propose a novel framework which consists of bi-audio-visual time-windows that span short video-clips labeled with discrete emotions. Attention is used to weigh these time-windows for multimodal learning and fusion. Experimental results on two datasets show that the proposed methodology can achieve an enhanced multimodal emotion recognition.

***Index Terms*—** attention, multimodal learning, audio-visual emotion recognition

1. INTRODUCTION

Emotions play a central role in human-human interaction [1]. They are highly sophisticated sub-conscious reactions, that are expressed through multiple cues, among which, the most prominent ones are visual and audio signals. In Affective Computing, AVER aims to efficiently capture these subtle emotional experiences and generate the proper actions, to have a natural Human-Computer Interaction (HCI) [2]. Applications of HCI can be found in entertainment [3], healthcare [4], and education [5].

Multimodal perception has shown a significant impact in terms of accurate performance [6]. However, this comes with challenges since there is not a linear relationship between their input and since each modality has distinct statistical properties [1]. In addition, in AVER, modalities' temporal dependencies and contribution to emotion perception are not fully exploited, as both modalities' importance varies over-time according to emotion classes [7]. For example, psychological studies show that the recognition speed of positive and negative emotions depends on the presentation of audio and video modalities [8].

Recently attention mechanisms have shown great success in a learning context in sequential data such as machine translation [9] and question answering [10]. This research aims to model and exploit the temporal relationship between audio-video cues utilizing a transformer-based self-attention mechanism. We propose a novel Multimodal Attention-mechanism for Temporal Emotion Recognition (MATER) framework to capture the audio-video inter-relationships. It is adopted for multimodal fusion and learning. We address the research question of how to efficiently utilize these signals over-time according to each modality's strength on emotions to maximize the automatic AVER performance.

MATER is a modality-specific framework, where learning is based on decision-level fusion which is performed on the prediction of each modality. This design allows the specialization of the framework to leverage the modality-specific properties in their data-stream. In this study, we investigate the benefit of attention mechanism for AVER. Besides, the model is extensively evaluated against several baselines and approaches such as Long-Short Term Memory (LSTM) and other state-of-the-art methods. We observe that multimodal recognition of emotions benefits from the attention mechanism.

Related Work: In multimodal context, attention has been applied for tasks such as Audio-Visual Speech Recognition AVSR [11], video captioning [12], and dialog systems [13]. For example, authors in [11] uses transformer architectures with Connectionist Temporal Classification (CTC) loss for recognizing phrases and sentences from audio and video signals. In [12], A self multimodal attention was used with LSTM to boost video captioning by learning from audio-video streams jointly. This approach exploited the multimodal input to generate coherent sentences.

In recent years, there has been a large body of work and interest in AVER using different approaches such as early and late fusion of different modalities [14, 15]. In addition, attention has been applied for emotion recognition. For example, authors in [16] utilized a self-attention mechanism to learn the alignment between text and audio for emotion recognition in speech. A self-attention layer was used to learn the alignment weights between speech frames and text words from different time-stamps. Authors in [14] proposed a recursive multi-attention with shared external memory based on Memory Network. Their cross-modal approach showed that gated memory effectively achieve multimodal emotion recognition.

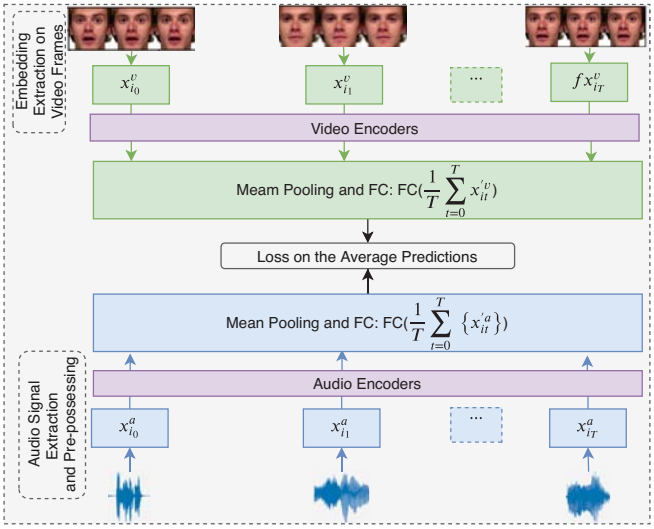


Fig. 1: The proposed methodology MATER for AVER. It has two data streams, composed of audio ($f^a(x^a)$) and video ($f^v(x^v)$) sub-networks.

2. METHODOLOGY

MATER as shown in Fig. 1, has two sub-networks, wherein each time-window (a sequence of frames), we employ the encoder part of the transformer [9] on the visual embeddings: X^v and another one on the audio embeddings: X^a. The novel bi-modal framework aims to study the temporal presentation of audio-visual cues for emotion recognition. The design of MATER is based on the following objectives and motivations:

- Emotion display consists of on-set, apex, and off-set phases, while the apex captures the maximum expressivity, thus, it is the segment considered in most research works [7]. Nevertheless, it is better not to pre-define these phases, since they depend on the emotions and the presented modalities. MATER is specialized in exploring and utilizing modalities' strength on these phases for better performance.
- Research demonstrated that emotion perception might require a different amount of time for an accurate detection [7]. Thus, these alterations could be exploited efficiently through a temporally-trimmed framework.

2.1. Input Modalities' Embeddings

In AVER, a dataset ($\mathbb{D}$) contains n short video clips with **a**udio and **v**isual (video) modalities, and each clip is annotated with a discrete emotion I^y:

$$\mathbb{D} = \{(x_1^v, x_1^a, I_1^y), (x_2^v, x_2^a, I_2^y), ..., (x_n^v, x_n^a, I_n^y)\}$$

where $x^{a,v}$ are the embeddings extracted from the audio or video raw-data. We took non-overlapping time-windows of 0.25 and 0.5 seconds as inputs for audio and visual models for embeddings extraction. These embeddings are then normalized with l_2-*normalization*.

2.1.1. Video Embeddings

In each time-window of a video clip, faces are detected and tracked using the Dlib library [17]. Subsequently, faces are cropped to 96×96 resolution. A pre-trained VGG-M model [18] is used to extract representations of a given face. We used the output from the final convolutional layer which has a 512-dimensional vector. As these representations are for each frame, we found out that mean-pooling through time-window frames' features have resulted in a good representation.

2.1.2. Audio Embeddings

We extract audio embeddings for a time-window using VGGish [19]. VGGish is a variant of VGG models, which was trained to generate high level and semantically useful embeddings for audio recordings. It was pre-trained with the YouTube-8M dataset [20], and we use the output of the last convolutional layer, which has 512-dimensional features. VGGish was trained with audio raw data using a 16 kHz mono sample rate. A spectrogram is computed using magnitudes of the Short-Time Fourier Transform (STFT) with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window [19]. In our case, as the time-windows length is either 0.25 or 0.5 seconds, the audio input size contains either 48×64 or 24×64 log mel spectrograms. These inputs were adapted to fit the requirements of the proposed MATER framework.

2.2. MATER's Components

We employ the encoder part of the transformer on each modality's embeddings. The encoder consists of a Multi-Head Self Attention (MHSA) layer and followed by an element-wise feed-forward layer. As suggested in the transformer [9], we also use 6 stacked encoder layers.

Positional Encoding (PE): the transformer adopts PE to make use of the order in a sequence and its time-information, instead of recurrence operations. It employs sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}})$$

where pos is the position and i is the dimension, which indicates that each position corresponds to a sinusoid [9]. The PEs are added to the embeddings prior to their flow to the encoder and they have the same dimensions (d_{model}) as the embeddings.

Multi-Head Self-Attention (MHSA): initially, the input of an encoder flows through a self-attention layer. In particular, scale dot-product attention is applied to the input of queries (Q), keys (K), and values (V). The attention function is applied on these packed matrices as:

$$Attention(Q, K, V) = softmax((\frac{QK^T}{\sqrt{d_k}})V) \quad (1)$$

where d_k dimensionality of K, and in the encoder of the self-attention layers it is always Q = K = V. Subsequently, multi-head attention is performed in parallel and their output is concatenated as follow:

$$MultiHead(X) = (concat(head_1, ..., head_h))W^O$$
$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Within MATER framework, as the input of each sub-network, we have audio-visual embeddings, X_t^m, where m refers to a modality: $m \in \{a, v\}$, and t represents a time-window: $t \in \{1, 2, ...T\}$. T is the maximum number of time-windows in a video clip. As a result, each sub-network of a modality has a sequence of embeddings $X^m = \{x_0^m, x_1^m, ..., x_T^m\}$ as input to the encoder, which attends to each time-window "token" in a different weight.

MHSA helps the model to learn representations from different subspaces at different positions. Following the MHSA layer, each output is fed onto the position-wise feedforward layer, independently for each time-window.

Prediction Layers: On the final output of the encoders, a mean pooling is applied for each modality separately:

$$X'^v = \frac{1}{T}\sum_{t=0}^{T} x_{it}'^v \text{ and } X'^a = \frac{1}{T}\sum_{t=0}^{T} x_{it}'^a \quad (2)$$

Two fully connected (FC) layers are applied on resulted audio (X^a) and video (X^v) representations as the prediction layers. The predictions from the two modalities are averaged and the network is optimized accordingly:

$$predictions = \frac{1}{2}\sum_{a,v}(W^m)^T X'^m + b^m \quad (3)$$

where W and b are the parameters of an FC.

3. EXPERIMENTS

The proposed framework's efficiency is evaluated on two public multimodal emotion recognition datasets, namely RAVDESS [21] and CREMA-D [22].

RAVDESS has two sets: speeches and songs subsets. We use the speech set as it is labeled with eight archetypal Ekmanian emotions [23]: anger, happiness, disgust, fear, surprise, sadness, calmness and neutral. The dataset has 24 subjects, 12 males and 12 females, with an age range of 21-33. It contains short speech video-clips of an average of 3.82 ± 0.34 seconds. The total number of videos is 1444.

CREMA-D consists of 7450 video clips for 91 subjects. The video-clips' average duration is 2.63 ± 0.53 seconds. Each video is labeled with six basic Ekmanian emotions: anger, disgust, fear, happiness, neutral, and sadness, with four different levels (intensities), low, medium, high and unspecified. The dataset includes people with a diverse background, in terms of gender, ethnicities, and ages.

Table 1: Evaluations' accuracies for various scenarios. RAVDESS and CREMA-D have an average of 3.82 ± 0.34, and 2.63 ± 0.53 seconds length video clips, respectively.

Dataset	#windows	Duration (seconds)	PE	MHSA	Accuracy %
RAVDESS	8	0.5	✓	✓	76.3
	8	0.5	✓	✗	70.6
	8	0.5	✗	✓	75.2
	8	0.5	✗	✗	69.4
	16	0.25	✓	✓	74.4
	16	0.25	✗	✗	66.2
CREMAD	6	0.5	✓	✓	67.2
	6	0.5	✓	✗	64.4
	6	0.5	✗	✓	65.0
	6	0.5	✗	✗	61.8
	12	0.25	✓	✓	66.4
	12	0.25	✗	✗	58.3

3.1. Training Details

MATER was optimized during the training phase using Adam optimizer [24], which is a variant of Stochastic Gradient Descent (SGD). Cross-entropy loss is used in this optimization. We use a batch size of 64 and the framework was trained for 300 epochs. Initially, the learning rate (lr) was set to $1e^{-6}$ and it was reduced if it reaches a plateau state after 20 epochs.

Evaluation Protocols: For both datasets, we use subject disjoint k-fold cross-validation. To have an equal number of subjects per fold, RAVDESS and CREMA-D were divided into 12 and 10 folds, respectively. In each fold, a subject's samples are either in a testing or a training fold. In addition, training and evaluations conducted separately on each dataset.

3.2. Baseline Models and Results

We examine the role of attention and the PE in audio-visual (AV) performance. A baseline model is introduced in which the attention and the PE are removed. The six stacked encoders' feedforward layers are kept which makes it a strong baseline as well. Keeping the depth of the models (including the baseline) the same provides a fair comparison. In addition, the flow of the embeddings, the training, and optimization processes are similar across the experiments. This baseline represents the case of averaging time-windows without weighing their importance for each modality. These studies aim to check the research's goal regarding the weighting mechanism that the attention scheme provides. In addition, it examines the role of PE in the framework.

These comparisons were tested on different numbers of time-windows. Due to different lengths of video-clips in CREMA-D and RAVDESS, the number of windows was set differently. We use sets of {8, 16} and {6, 12} time-windows for RAVDESS and CREMA-D, respectively. As shown in Table 1, we notice that the best performance on both datasets is achieved when using MATER with the PE and the attention, where the accuracy reaches 76.3% and 67.2% for RAVDESS

Table 2: AV accuracies of MATER and other related work.

Approach	CREMA-D	RAVDESS
Human Perception: AV	63.6	80.0
Dual Attention with LSTM: AV [14]	65.0	58.3
Metric Learning for AVER [25]	66.5	(not available)
MATER: $AV_{+PE+MHSA}$	67.2	76.3

and CREMA-D, respectively. PE enhances the performance since it gives the system the time and the order information, where the improvement over using only the attention is at least 1%. This information is further utilized through the MHSA. Moreover, PE's impact is more obvious when the number of time-windows is large.

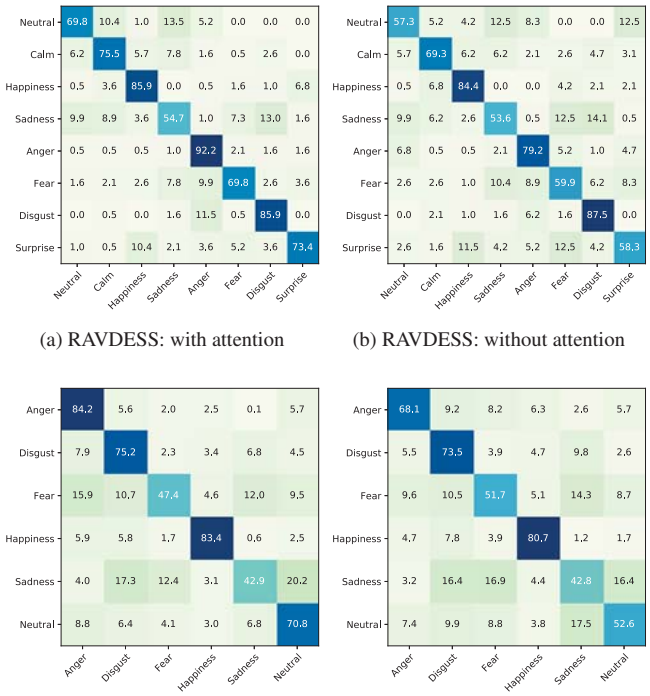
In the baseline results of the framework, in case of not using both the PE and the attention, we observe that the performance drop by at least 5% and 3% for RAVDESS and CREMA-D, respectively. This gap increases when the number of time-windows is doubled, where the improvement reaches at least 8%.

Comparisons to other methods: previous work results in both datasets, including human performance, are presented in Table 2. MATER results in this table are obtained using the attention, with 8, and 6 time-windows for CREMA-D and RAVDESS, respectively. Our approach outperformed both human-perception and the recently published results in [14, 25]. In [14], the performance (65.0% and $58.3.\%$ accuracies) was obtained by combining facial and audio temporal features with LSTM using Dual-Attention. In [25], a metric learning approach was applied to fuse audio-video modalities. MATER's results show its efficiency for enhanced joint multimodal learning and fusion. Another reason behind these improvements is that MATER deals with the interaction of the multi-modal data over-time using the time-windows segments. This makes the framework weighs and evaluates the importance of the two modalities per emotion across time.

Confusion Matrices (CMs). CMs displayed in Fig. 2 show the achieved performance of our approach on RAVDESS and CREMA-D classes and the degree in which each emotion was confused to the other ones. The x-axis represents the intended emotions, and the y-axis shows the predicted emotions. Without exception, the diagonal elements have the highest accuracies, which indicates the high classification accuracy of the intended emotions. More importantly, the improvement margin over the baseline is more obvious in emotions such as anger and neutral.

In terms of MATER performance on emotions, anger, neutral, disgust, and happiness have higher accuracy detection, compared to fear and sadness. While we notice that, e.g., fear and sadness were confused with other emotions in varied ratios. These results are also compatible with the reported human perception and confusion in [22, 21].

Ablation study. Table 3 introduces the accuracies of the underlying Audio (A) and Video (V) modalities, within the



(a) RAVDESS: with attention (b) RAVDESS: without attention

(c) CREMA-D: with attention (d) CREMA-D: without attention

Fig. 2: CM between true and predicted labels.

Table 3: Multimodal and individual performance of MATER with and without attention.

Dataset	Attention	A	V	AV
RAVDESS	✓	59.2	58.2	76.3
RAVDESS	✗	60.7	56.0	69.4
CREMA-D	✓	57.5	51.7	67.2
CREMA-D	✗	56.0	49.0	61.8

framework. These results show the sub-modalities' contribution in the performance of the framework. They show that the accuracy is significantly increased when using both audio and video modalities. The improvement is at least 10% over the uni-modal perception. Notably, attention helps in the multimodal fusion due to the weighting mechanism of the modalities over-time. This highlights the essential role of multimodal perception to obtain enhanced emotion recognition.

4. CONCLUSION

This research highlights the importance of exploiting audio-video signals' temporal strength for emotion recognition. We utilize the attention mechanism on audio-visual embeddings over time-windows to leverage their properties for emotion recognition. Evaluation of two datasets shows that the proposed method with the attention mechanism improves the performance over the baseline significantly. They show the advantage of weighing the contribution of each modality.

5. REFERENCES

[1] Philipp V Rouast, Marc Adam, and Raymond Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019.

[2] Rosalind W Picard, *Affective computing*, MIT press, 2000.

[3] Sarah Cosentino et al., "Group emotion recognition strategies for entertainment robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 813–818.

[4] Leandro Y Mano et al., "Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition," *Computer Communications*, vol. 89, pp. 178–190, 2016.

[5] Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[7] Yelin Kim and Emily Mower Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 92–99.

[8] Pashiera Barkhuysen, Emiel Krahmer, and Marc Swerts, "Crossmodal and incremental perception of audiovisual cues to emotional speech," *Language and speech*, vol. 53, no. 1, pp. 3–30, 2010.

[9] Ashish Vaswani et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Triantafyllos Afouras et al., "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[12] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei, "Learning multimodal attention lstm networks for video captioning," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 537–545.

[13] Chiori Hori et al., "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2352–2356.

[14] Rory Beard et al., "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018, pp. 251–259.

[15] Samira Ebrahimi Kahou et al., *Combining modality specific deep neural networks for emotion recognition in video*, ACM, 2013.

[16] Haiyang Xu et al., "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[17] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[18] Samuel Albanie and Andrea Vedaldi, "Learning grimaces by watching tv," *BMVC*, 2016.

[19] Shawn Hershey et al., "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[20] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[21] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[22] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[23] Paul Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[25] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis, "Metric learning based multimodal audio-visual emotion recognition," *IEEE MultiMedia*, 2019.