Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13557-w

1226: DEEP-PATTERNS EMOTION RECOGNITION IN THE WILD

Joint modelling of audio-visual cues using attention mechanisms for emotion recognition

Esam Ghaleb[1] · Jan Niehues[1] · Stylianos Asteriadis[1]

Received: 2 December 2021 / Revised: 11 April 2022 / Accepted: 14 July 2022

Abstract

Emotions play a crucial role in human-human communications with complex socio-psychological nature. In order to enhance emotion communication in human-computer interaction, this paper studies emotion recognition from audio and visual signals in video clips, utilizing facial expressions and vocal utterances. Thereby, the study aims to exploit temporal information of audio-visual cues and detect their informative time segments. Attention mechanisms are used to exploit the importance of each modality over time. We propose a novel framework that consists of bi-modal time windows spanning short video clips labeled with discrete emotions. The framework employs two networks, with each one being dedicated to one modality. As input to a modality-specific network, we consider a time-dependent signal deriving from the embeddings of the video and audio modalities. We employ the encoder part of the Transformer on the visual embeddings and another one on the audio embeddings. The research in this paper introduces detailed studies and meta-analysis findings, linking the outputs of our proposition to research from psychology. Specifically, it presents a framework to understand underlying principles of emotion recognition as functions of three separate setups in terms of modalities: audio only, video only, and the fusion of audio and video. Experimental results on two datasets show that the proposed framework achieves improved accuracy in emotion recognition, compared to state-of-the-art techniques and baseline methods not using attention mechanisms. The proposed method improves the results over baseline methods by at least 5.4%. Our experiments show that attention mechanisms reduce the gap between the entropies of unimodal predictions, which increases the bimodal predictions' certainty and, therefore, improves the bimodal recognition rates. Furthermore, evaluations with noisy data in different scenarios are presented during the training and testing processes to check the framework's consistency and the attention mechanism's behavior. The results demonstrate that attention mechanisms increase the framework's robustness when exposed to similar conditions during the training and the testing phases. Finally, we present comprehensive evaluations of emotion recognition as a function of time. The study shows that the middle time segments of a video clip are essential in the case of using audio modality. However, in the case of video modality, the importance of time windows is distributed equally.

✉ Esam Ghaleb
esam.ghaleb@maastrichtuniversity.nl

Extended author information available on the last page of the article.

Published online: 05 August 2022

Keywords Affective computing · Emotion recognition · Multimodal learning · Attention mechanisms

1 Introduction

Emotions are expressed through multiple cues. Among these, the most prominent ones are visual and audio signals. Emotion-related cues are usually complementary to each other. Observing both visual data (e.g. facial expression), along with voice characteristics in audio (e.g. prosodics, voice frequential components, or deep features) can help in an overall improvement of emotion recognition [7]. Recently, multimodal emotion recognition has gained a notable amount of research [15, 33, 34]. Practically, it is part of Affective Computing (AC), which is a multidisciplinary field that aims to automate the process of emotion recognition, simulation, and inducement. AC is conceptually grounded in affective science. The foundations of AC are in the fields of psychology, neurology, and sociology. Kappas et al. in [14, 23] suggested that affective states are encompassing neurological changes, physiological responses, body expressions, and cognitive and metacognitive states. Besides, these changes are modulated by contextual and social influences. Emotion-related cues are usually complementary to each other and observing both visual data (e.g. facial expression), along with voice characteristics in audio (e.g. prosodics, voice frequential components, or deep features) can help in an overall improvement of emotion perception and recognition [7].

However, multimodal fusion comes with challenges since there is not a linear relationship between modalities' raw data and since each modality has distinct statistical properties [34]. Therefore, in this work, we propose a dedicated solution to address the challenges of multimodal fusion in the context of emotion recognition from audiovisual cues. Emotion recognition using multiple data sources is an essential factor to increase the quality of communications in human-computer interaction. For example, recent technological advancements brought interactivity between people and digital devices to a completely different level, making computers and mobile phones an important part of our daily lives. Hence, automatic systems with affective capabilities can be essential in many applications of affective computing, which range from entertainment [13], healthcare [28], and education [38].

Uni-modal perception of these signals is usually complementary to each other and can improve the overall observation of phenomena such as emotions. In Affective Computing, Audio-Video Emotion Recognition (AVER) aims to efficiently capture these subtle emotional experiences and generate the proper actions, to have a natural Human-Computer Interaction (HCI) [30]. In addition, in AVER, modalities' temporal dependencies and contribution to emotion perception are not fully exploited, even though audio-video modalities' importance varies over time according to the expressed and perceived emotions [25]. For example, a number of studies coming from the field of psychology have demonstrated that positive and negative emotions can be recognized at an early or late stage during the expression, depending on the available modalities [8]. Moreover, various studies on human ratings for emotion through audio-visual cues revealed interesting observations in terms of which modality is more useful for which kind of emotions. In [10], the authors showed that the recognition of disgust and fear is better with audio-visual cues. On the other hand, anger and happiness are recognized accurately with single modalities. These observations are also consistent with our findings throughout this dissertation. For example, researchers in [10] showed that the human visual perception alone achieved a 69.0% accuracy, while the perception in audio alone was 45.5%. However, the presentation of both visual and auditory signals to human raters increased their perception by at least 5.8%. From a computational perspective, D'Mello and Kory conducted a meta-analysis on multimodal emotion recognition systems. Their study revealed that multimodal emotion detectors are consistently better than their underlying unimodal detectors, with an average improvement of 9.8%.

This study focuses further on attending to the informative time segments in audio and visual cues for multimodal emotion recognition. It addresses the following research question: *how can we capture the contributions of the temporal dynamics of affect display using attention mechanisms?* A large body of research has recently shown that attention mechanisms result in great success when modelling and interpreting sequential data, with applications in machine translation [36] and natural human-machine communications [16] (e.g. chatbots). For this purpose, we utilize the Transformer's self-attention mechanisms [36]. The Transformer is currently the state-of-the-art approach for many tasks with sequential data. We, thus, propose a novel Multimodal Emotion Recognition Metric Learning (MATER) framework, adapted to the needs of multimodal fusion across time windows of audiovisual cues. The framework is introduced to efficiently utilize these cues over time according to each modality's strength on emotions to maximize the automatic AVER performance. Furthermore, MATER is a modality-specific framework, where learning is based on decision-level fusion. This design allows the specialization of the framework to leverage modality-specific properties in their data stream. The contributions of this research can be summarized as follows:

- Utilizing attention mechanisms, we propose a novel methodology to address audiovisual emotion recognition over time. The methodology is comprehensively evaluated against several baselines and approaches such as Long-Short Term Memory (LSTM). Based on our experimental findings, we conclude that employing attention mechanisms benefit computational models, as was our initial intuition.
- Our work presents a framework and experimental evaluations to understand underlying principles of emotion recognition as functions of three separate setups in terms of modalities: audio only, video only, and the fusion of audio and video, linking the outputs of our propositions to research from psychology.
- We study the impact of attention mechanisms within the proposed framework on (1) the bimodal emotion recognition, (2) the performance under challenging conditions (i.e., when using noisy data), and (3) the perception of emotions over time through audiovisual cues.

This paper is organized as follows. Section 2 introduces a literature review on emotion recognition and how attention mechanisms are applied in this field. Section 3 explains the proposed methodology. Section 4 presents experimental evaluations of the proposed approach, studying the impact of attention mechanisms on bimodal emotion recognition. Section 5 elaborates the re-training strategies of MATER using noisy data and investigates the role of attention mechanisms in increasing the framework robustness under challenging conditions. Section 6 gives an extended evaluation of the framework to study the role of time in emotion recognition. Finally, Section 7 concludes the research and highlights its findings.

2 Related work

2.1 Human perception of audio-visual cues

A large body of work in psychology shows the importance of multimodal perception for emotion recognition. For example, Auberge et al. [5] show that even in visible emotions (e.g. amusement), audio modality carries important information that is related to a smiling face. Their study proves that acoustic information clearly interacts with visual decoding. Besides, Barkhuysen et al. [8] studied how visual cues from the speaker's face relate to emotions. The study found out that positive emotions can be detected accurately with visual information, while negative emotions are better perceived using only the audio modality. However, audio-visual modalities usually increase perception accuracy. Similarly, automatic emotion recognition should be capable of factoring affective states when having a multimodal presentation [18]. In addition, the human rating of emotions revealed interesting observations in terms of which modality is more influential for which kind of emotion. In [10], Cao et al. showed that accurate recognition of disgust and fear requires audio-visual cues, while anger and happiness are recognized based on single modalities. Moreover, Cao et al. found that the human visual perception alone achieved a 58.2% accuracy, while the perception in audio alone dropped to 40% accuracy. However, the presentation of both visual and auditory signals to human raters improved their perception accuracy by at least 5%.

2.2 Multimodal emotion recognition

In automatic emotion recognition, many works initially focused on individual modalities (e.g. emotion recognition using facial expressions, speech, or data from wearable sensors). However, the focus soon shifted toward Multimodal Emotion Recognition (MER) [14, 33, 34]. MER is more realistic and resembles the way humans detect emotions. Nonetheless, this is a challenging task for several reasons. Firstly, emotions have a highly complex nature, making it difficult to model and frame them for even humans themselves. Nevertheless, researchers have established a number of theories to simplify these sophisticated experiences [17, 31]. Secondly, multimodal data come from varying numbers of sensors and have different properties, and distributions [29]. Therefore, simple fusion or learning algorithms might not be useful to capture the dependencies and complementary information between these modalities due to the non-linear relationship between them. As a result, there is a need to develop systems to deal with these challenges for accurate and better multimodal learning schemes.

In this work, we propose joint modelling of audio-visual cues based on time windows. The idea is to utilize temporal and multimodal information to enhance emotion recognition. As will be demonstrated through our experiments (Section 4), the representations (deep embeddings) involved, as well as the architectural characteristics of our approach, constitute the proposed methodology a promising track towards multimodal emotion recognition.

2.3 Attention mechanisms for multimodal learning

In a multimodal context, attention mechanisms have been applied for tasks such as Audio-Visual Speech Recognition (AVSR) [2], video captioning [40], and dialogue systems [22]. For example, authors in [2] used transformer architectures with Connectionist Temporal

Classification (CTC) loss for recognizing phrases and sentences from audio and video signals. In [40], self multimodal attention was used with LSTMs to boost video captioning by learning from audio-video streams jointly. This approach exploited the multimodal input to generate coherent sentences.

In addition, attention mechanisms have been applied for emotion recognition. For example, authors in [41] 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. In addition, Wu et al., in [39], employed transformer-based self-attention to attend the emotional autobiographical narratives. In their study, attention mechanisms were found to be powerful in a combination of Memory Fusion Network for multimodal fusion of audio, video, and text modalities. Authors in [9] proposed a recursive multi-attention with shared external memory based on Memory Networks. Their cross-modal approach showed that gated memory could achieve robust results in multimodal emotion recognition. Our work addresses emotion expression as a function of time windows and models the joint learning of audio-visual cues using attention mechanisms. Our method succeeds in enhancing multimodal recognition and offers a framework to understand the behaviour of temporal audio-video emotion recognition and the benefit of their joint modelling.

3 The proposed methodology

This section describes the main components of the proposed methods: namely, we provide details regarding the extraction of audio-visual embeddings, application of the Transformer attention mechanisms on the embeddings of time windows in a video clip, and their joint multimodal fusion.

Multimodal Attention mechanisms for Temporal Emotion Recognition (MATER), shown in Fig. 1, consists of two networks, with each one being dedicated to one modality. As input to a modality-specific network, we consider a time-dependent signal deriving from the embeddings of the video ($X^{(v)}$) and audio ($X^{(a)}$) modalities. We employ the encoder part of the Transformer [36] on the visual embeddings, $X^{(v)}$, and another one on the audio embeddings, $X^{(a)}$. The embeddings are obtained from VGG models, and the applied audio and video encoders have the same architecture. As motivated in [35, 37], multimodal deep learning is a challenging task, due to the increased capacities of Deep Neural Networks (DNNs), in the case of more than one modality exists. For this reason, in this study, the architecture is based on a late joint fusion to avoid overfitting one modality and allow the two sub-networks to generalize at different rates. In addition, from an analytical point of view, the design of MATER is based on the following 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 [25]. 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' correlation with emotions on these phases for robust performance.
- Research has demonstrated that emotion perception might require a different amount of time for an accurate detection [25], depending on the expressed emotion and involved modalities. Thus, these alterations could be exploited efficiently through a temporal framework.

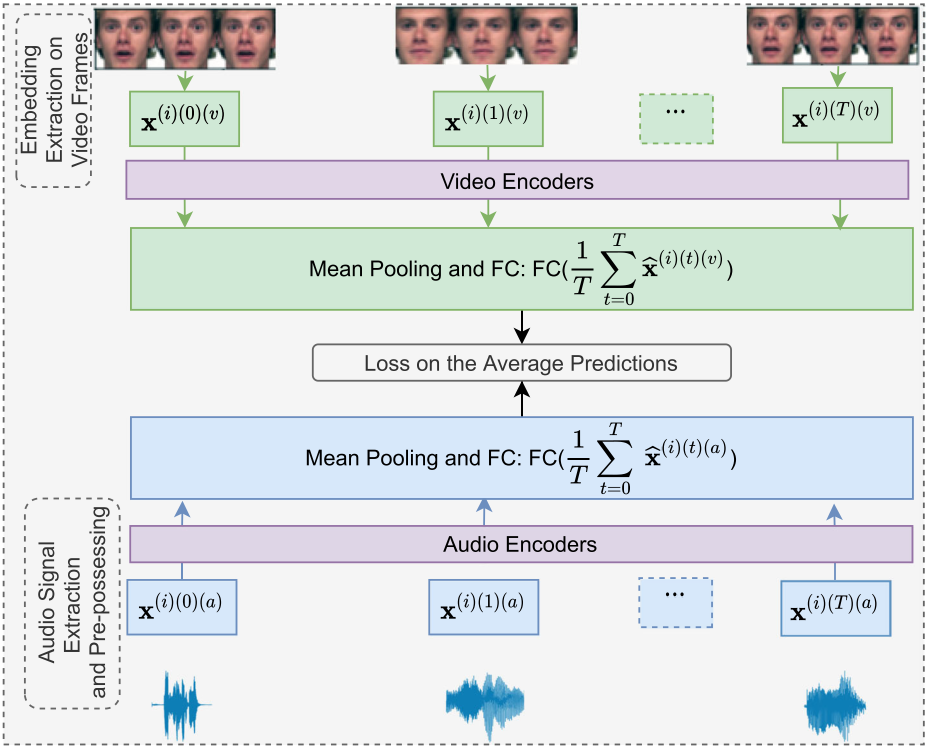

Embedding Extraction on Video Frames
$\mathbf{x}^{(i)(0)(v)}$
$\mathbf{x}^{(i)(1)(v)}$
...
$\mathbf{x}^{(i)(T)(v)}$
Video Encoders
Mean Pooling and FC: $FC(\frac{1}{T}\sum_{t=0}^{T}\hat{\mathbf{x}}^{(i)(t)(v)})$
Loss on the Average Predictions
Mean Pooling and FC: $FC(\frac{1}{T}\sum_{t=0}^{T}\hat{\mathbf{x}}^{(i)(t)(a)})$
Audio Encoders
$\mathbf{x}^{(i)(0)(a)}$
$\mathbf{x}^{(i)(1)(a)}$
...
$\mathbf{x}^{(i)(T)(a)}$
Audio Signal Extraction and Pre-possessing

Fig. 1 An illustration of the proposed framework MATER for AVER An illustration of the proposed framework, MATER, for AVER. It has two data streams, composed of two sub-networks, applied on raw audio ($f^{(a)(i)}(x^{(a)(i)})$) and video ($f^{(v)}(x^{(v)})$) data coming from a video clip, i

3.1 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 c,

$$\mathbb{D} = \{(\boldsymbol{x}^{(v)(1)}, \boldsymbol{x}^{(a)(1)}, c^{(1)}), (\boldsymbol{x}^{(v)(2)}, \boldsymbol{x}^{(a)(2)}, c^{(2)}), ..., (\boldsymbol{x}^{(v)(n)}, \boldsymbol{x}^{(a)(n)}, c^{(n)})\},$$

where $\boldsymbol{x}^{(a,v)}$ are the embeddings extracted from the audio or video raw-data. In this work, we consider non-overlapping time windows of 0.25 and 0.5 seconds, as inputs to the audio and visual models for embeddings extraction. These embeddings are then normalized with l_2-*normalization* to have zero mean and unit length.

3.1.1 Video embeddings

In each time window of a video clip, faces are detected and tracked using Ensemble of Regression Trees (ERT) method (introduced in [24]). Subsequently, faces are cropped to 96 × 96 resolution. A pre-trained VGG-M model [3, 12] on the Facial Emotion Recognition (FER) dataset [20] is used to extract representations of a given facial image. We used the output from the final convolutional layer, which corresponds to a 512-dimensional vector. As these representations are for each frame, we found that mean-pooling through the features of time windows' frames gives robust representations. Another alternative pooling scheme can be max-pooling, however, we observed that this scheme was inferior to the adopted one.

3.1.2 Audio embeddings

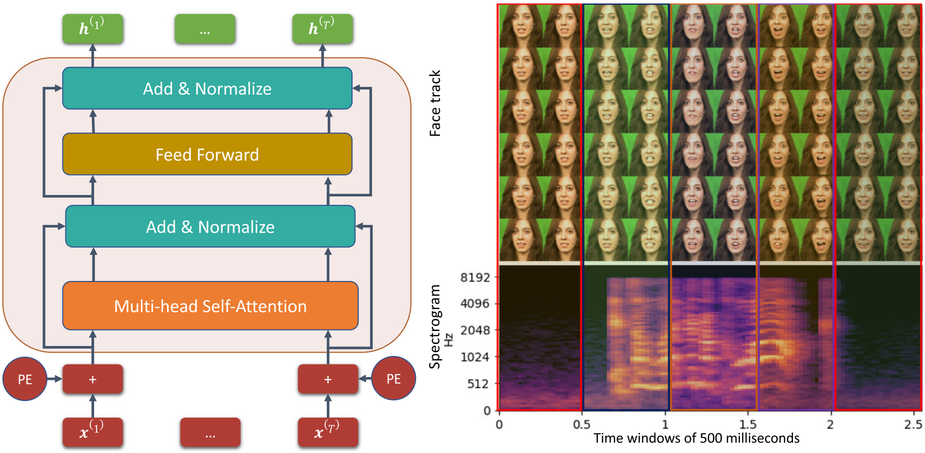
We extract audio embeddings for a time window using VGGish [21]. VGGish is a variant of VGG models, which was trained to generate high level and semantically useful embeddings for audio recordings. It was pre-trained with the YouTube-8M dataset [1], and we use the output of the last convolutional layer, which corresponds to a 512-dimensional vector. VGGish was trained with audio data using a 16 kHz mono sample rate. Specifically, 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 [21]. In our case, as the time windows length is either 0.25 or 0.5 seconds, the audio input size contains either 24 × 64 or 48 × 64 log mel spectrograms. Each example covers 64 mel bands and 48 or 24 frames of 10 ms each. These inputs were adapted to fit the requirements of the proposed framework.

3.2 Framework's components

MATER employs the attention mechanisms of the Transformer, which was introduced in [36]. The Transformer is a neural network architecture that has an encoder-decoder structure.

3.2.1 Encoder

In our study, we employ the encoder part of the Transformer on each modality's time segments. Figure 2a shows the architecture of an encoder in the Transformer. As shown in



(a)The encoder part of the Transformer is used for each modality. Note that, in our framework, we stack 6 encoder layers.

(b)An example of non-overlapping time windows of 500 milliseconds.

Fig. 2 Transform's Encoder and an example of non-overlapping time windows

the detailed Fig. 2a, an encoder consists of a Multi-Head Self Attention (MHSA) sublayer and is followed by an element-wise fully connected feed-forward sublayer. In addition, a residual connection between the two sublayers is employed and followed by layer normalization. As a result, according to the authors of [36]'s terminology, the output of each layer is "$LayerNorm(\boldsymbol{x} + Sublayer(\boldsymbol{x}))$", where $Sublayer(\boldsymbol{x})$ refers to the function of the MHSA or the element-wise feed-forward sublayers, and $\boldsymbol{x}$ is time-dependent embeddings (feature vector). Finally, all sublayers in the encoder produce outputs with the same dimensions, e.g. $d = 512$, to facilitate the residual connections. The number of stacked encoders could vary, and in the original paper [36] it was set to 6. Before feeding the encoder blocks with the sequential data, a positional encoding operation is applied by adding time information to the input embeddings.

3.2.2 Audio-video inputs

As input to each sub-network (audio and visual encoders), we consider audio-visual embeddings, $X^{(m)(t)}$, where m refers to a modality: $m \in \{a, v\}$, and t represents a time window: $t \in \{1, 2, ...T\}$. T is the number of time windows in a video clip (as shown in Fig. 2b). As a result, each sub-network of a modality has a sequence of embeddings,

$$X^{(m)} = \{\boldsymbol{x}^{(m)(0)}, \boldsymbol{x}^{(m)(1)}, ..., \boldsymbol{x}^{(m)(T)}\},$$

as an input to its encoders, which attends to each time window "token" with a different weight.

3.2.3 Positional encoding

Prior to feeding the audio-visual embeddings to the encoders, positional encoding is applied on the embeddings. The Transformer does not make use of recurrence or convolutional operations; rather, it adopts Positional Encoding (PE) in order to make use of ordinal information in a sequence. In particular, "positional encodings" are added to the sequential input of the encoder, e.g., in our case, the embeddings of audio-visual time windows of a video clip. This addition (sum) operation is applied once, before the flow of the audio-visual inputs to the encoders. Besides, they have the same dimensions (d) as the input embeddings to facilitate their sum. The authors of the Transformer [36] proposed to employ PE using sine and cosine functions, which are fixed ones, with variant frequencies as follows:

$$\begin{aligned} pe_{(t,2i)} &= \sin\left(t/10000^{2i/d}\right) \\ pe_{(t,2i+1)} &= \cos\left(t/10000^{2i/d}\right), \end{aligned} \tag{1}$$

where t indicates a time window and i refers to a specific dimension in the embeddings of this time window.

3.2.4 Multi-Head Self-Attention (MHSA)

The input of an encoder flows through a self-attention layer. The self-attention employed in the Transformer is used to assign a weighing score for each token (time window) in a time series. In our case, the tokens are the given embeddings of each time window in each modality within the proposed framework. In a video clip, self-attention focuses on specific time windows where emotion expression is strong by automatically assigning activations (weights) to these time windows. In addition, attention mechanisms help DNN models to learn context related to time and proximity of sequential inputs, e.g. the audio-video time windows of a video-clip: $[\boldsymbol{x}^{(m)(1)}, ..., \boldsymbol{x}^{(m)(T)}]$. A self-attention layer aims to weigh these vectors with respect to each other and results in the following weighted outputs: $[\boldsymbol{h}^{(m)(1)}, ..., \boldsymbol{h}^{(m)(T)}]$, where, e.g., $\boldsymbol{h}^{(m)(2)}$ is a weighted vector over all the input sequence. For instance, as shown in Fig. 2b, the embeddings of the facial expressions in the second time window can be associated with the ones in the middle time windows due to their proximity.

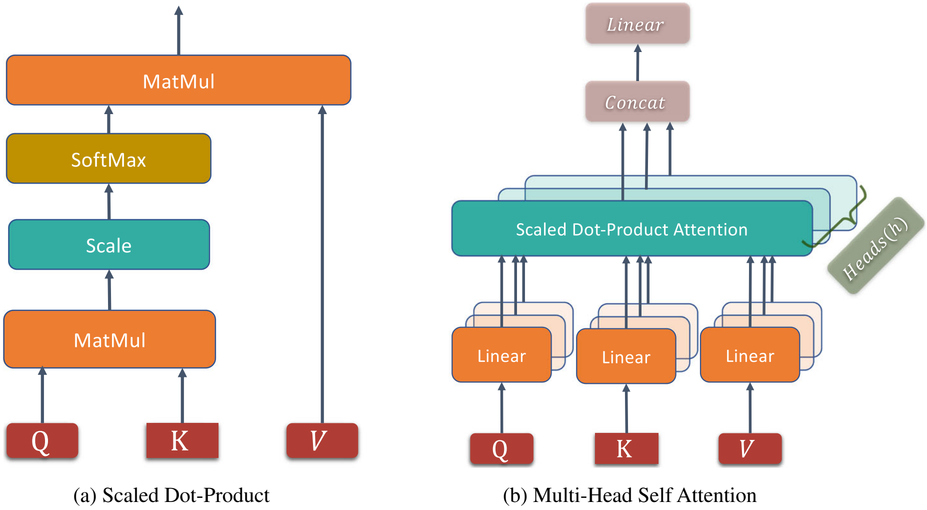
Mathematically, Vaswani et al. in [36] proposed using the "scaled dot-product attention", which is illustrated in Fig. 3a and formulated as follows:

$$Attention(Q^{(m)}, K^{(m)}, V^{(m)}) = softmax(\frac{Q^{(m)}K^{(m)T}}{\sqrt{d_k}})V^{(m)} \tag{2}$$

where queries ($Q^{(m)}$), keys ($K^{(m)}$), and values ($V^{(m)}$) matrices are created from the same input in a sequence. This is since the encoder part of the Transformer employs a self-attention mechanism by attending to its input sequence, $X^{(m)}$. In addition, since we employ two encoders for audio and video modalities separately, the scaled dot-product attention is applied on each modality accordingly.

Semantically, we can use the concepts of queries, keys, and values from information retrieval to explain the computation of attention mechanisms. In particular, the computation of attention can be considered as mapping a set of target vectors (*queries*) with a set of candidate vectors (*keys*). Subsequently, the scores resulting from these mappings are used to compute the weighted combination of the *values*, where the scores indicate the compatibility (similarity) of each *key* with the *query*. In our case, a query can be embeddings at the t time window. At the same time, the set of keys and the values are the whole sequences of the modalities embeddings (all the time windows). Hence, all queries ($Q^{(m)}$), keys ($K^{(m)}$), and values ($V^{(m)}$) matrices are created from the same sequence.

Moreover, MHSA is a key component in the Transformer architecture. As illustrated in Fig. 2a, following the addition of the PEs to the audio and video sequence embeddings, the resulting embeddings are fed forward through the MHSA layer. Specifically, MHSA splits the learning loads to learn context information over several heads. In particular, for



(a) Scaled Dot-Product

(b) Multi-Head Self Attention

Fig. 3 The operations within the MHSA [36]

the queries, keys, and values, we learn linear projections with d_q, d_k d_v dimensions, respectively. In the encoder part of the Transformer, these dimensions are the same. For example, in our case, each time window's embeddings in a video clip, $\boldsymbol{x}^{(m)(t)}$, has 512 dimensions (i.e. $d = 512$), and we use 8 attention heads. As a result, in a head (i), the linear projection matrices are as follows: $W_i^{(q)(m)}$, $W_i^{(k)(m)}$, and $W_i^{(v)(m)} \in \mathbb{R}^{d_k \times d}$.

Practically, for each modality, MHSA is applied on the input of queries ($Q^{(m)}$), keys ($K^{(m)}$), and values ($V^{(m)}$), shown in Fig. 3b. Subsequently, the resulting outputs from different attention heads, with d_k dimensions each, are concatenated and projected again (with $W^{(o)(m)}$ linear projection) to obtain the final weighted matrices. These matrices are used in the following sublayer, namely: element-wise feedforward sublayer. Mathematically, MHSA computations are performed as follows:

$$\begin{aligned} MHSA(X) &= W^{(o)(m)}(concatenate(head_1^{(m)}, ..., head_h^{(m)})), \\ \text{where } head_i^{(m)} &= Attention(W_i^{(q)(m)} Q^{(m)}, W_i^{(k)(m)} K^{(m)}, W_i^{(v)(m)} V^{(m)}), \end{aligned} \tag{3}$$

where $W_i^{(q)(m)}$, $W_i^{(k)(m)}$, and $W_i^{(v)(m)}$ are learnable linear transformations that help the self-attention mechanisms to get stronger representations and exploit the context in a given sequence of audio-video time windows.

3.2.5 Fusion: prediction layers

On the final output of the last modalities' encoder layers, hidden representations $\boldsymbol{h}^{(t)}$ are obtained for each time window. In our framework, the last encoder layer is the sixth encoder layer since we used six encoder layers for each modality. Subsequently, over the input sequence T, we apply a mean pooling for each modality, separately, in order to get the final audio and video representations:

$$\hat{\boldsymbol{h}}^{(v)} = \frac{1}{T} \sum_{t=0}^{T} \boldsymbol{h}^{(v)(t)} \text{ and } \hat{\boldsymbol{h}}^{(a)} = \frac{1}{T} \sum_{t=0}^{T} \boldsymbol{h}^{(a)(t)}. \tag{4}$$

Two fully connected (FC) layers are then applied on the resulting audio ($\hat{\boldsymbol{h}}^{(a)}$) and video ($\hat{\boldsymbol{h}}^{(v)}$) representations as the prediction layers. The predictions from the two modalities are averaged, and the network is optimized accordingly,

$$predictions = \frac{1}{2} \sum_{m \in \{a,v\}} (W^{(m)})\hat{\boldsymbol{h}}^{(m)} + \boldsymbol{b}^{(m)}, \tag{5}$$

where $W^{(m)}$ and $\boldsymbol{b}^{(\boldsymbol{m})}$ are the parameters of a fully connected layer. These averaged predictions are normalized via softmax operation and are used to compute the cross-entropy loss.

4 Experimental results

This section presents the experimental setup, implementation details, and the general evaluation metrics and results of MATER. The proposed framework's efficiency is evaluated on two public multimodal emotion recognition datasets, namely Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [27] and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [10].

RAVDESS [27] is an audio-visual dataset of dynamic expressions for basic emotions. It contains a large number of songs and speech recordings, each available in audio-only, video-only, and audio-visual formats. Moreover, 247 individuals from North America provided ratings. Participants with high reliability rigorously validated the dataset's emotions. Each clip was rated 10 times on emotional intensity, validity, and genuineness. High levels of emotional validity, inter-rater reliability, and test-retest intra-rater reliability were reported. RAVDESS is a gender-balanced dataset of 24 actors who performed vocalizations with emotions that include: anger, calmness, disgust, fear, happiness, sadness, and surprise. Actors also performed neutral vocalizations. In this paper, we chose to use the speech part of the dataset as it is labelled with eight archetypal emotions. This subset contains a total of 1444 recordings. The recordings in RAVDESS have an average duration of 3.82 ± 0.34 seconds. Raters' perception was reported to be: 62.0%, 72.0%, and 80.0%, for audio-only, video-only, and audio-video modalities, respectively.

CREMA-D [10] is an audio-video emotion expression dataset. It contains 7442 clips from 91 actors (43 females and 48 males). Participants' age ranges between 20 and 74, and they come from a variety of races and ethnicities, i.e. Asian, African American, Caucasian, and Hispanic. Actors were asked to speak 12 sentences in five different emotions, namely, anger, disgust, fear, happiness, and sadness, or in neutral. In CREMA-D, video clips have an average length of 2.63 ± 0.53 seconds. Furthermore, authors of the CREMA-D dataset asked 2443 participants to rate the emotions and their intensities on three settings: video alone, audio alone, and full audio-video clips. Each participant rated 90 clips (i.e. 30 audio, 30 visual, and 30 audio-visual). 95% of the video-clips have at least 8 ratings. An extensive evaluation was then provided on their responses. We report, here, the recognition rates that are based on the relative majority. The relative majority (i.e. a plurality) is measured when an emotion gets the largest share of the votes (ratings) in comparison with the rest of the other emotions. Therefore, this emotion is labeled as the perceived emotion. In this case, the recognition rates for audio-only, video-only, and bimodal audio-video perception are 45.5%, 69.0%, and 74.8%, respectively.

4.1 Training details

MATER was optimized during the training phase using Adam optimizer [26], 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 reached 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 testing or a training fold. In addition, training and evaluations have been conducted separately on each dataset. The RAVDESS and CREMA-D datasets provide extensive studies on human perception of emotions, forming a benchmark and a reference for our evaluations. The analysis of human perception is useful for our studies since this work focuses on spotting the temporal dynamics of emotions within short video clips on providing a meta-analysis of automatic perception of emotions.

Finally, in our experiments, we take into consideration several evaluation criteria: (i) Accuracy, which is the number of correctly classified video samples (ii) Confusion Matrix between the ground truth and the predicted emotion labels and (iii) KL-Divergence and entropy differences between audio and video predictions.

4.2 Baseline models and results

To evaluate MATER, we built baseline models for analytical comparisons, which examine the role of attention mechanisms in audio-visual (AV) emotion recognition. MATER consists of time windows based audio-visual embeddings, 6 stacked audio-visual encoders (each one has positional encoding, multi-head self-attention, and feedforward layers with their residual connections). To check the impact of MATER's components, MHSA, PE, or both are removed from the baseline models. The baseline and attention models (where MHSA and PE are kept) have the same number of layers and use the same settings in terms of audio-visual embeddings. The six stacked encoders' feedforward layers are kept, making it a strong baseline and a fair comparison.

The comparisons presented in Table 1 aim to check the research's goal regarding the weighing that the attention mechanisms scheme provides. It also examines the role of PE in the framework, where PEs are added to the embeddings. Due to different lengths of video clips in CREMA-D and RAVDESS, the number of windows was set differently. We use sets of f8, 16g and f6, 12g time windows for RAVDESS 14 Joint Modelling of Audio-visual Cues for Emotion Recognition and CREMA-D, respectively. As shown in column 8 of Table 1, the best AudioVisual (AV) performance on both datasets is achieved when using MATER with PE and attention (provided through MHSA), where the accuracy reaches

Table 1 This table presents a detailed performance analysis using different parameters, such as MHSA and PE, and windows' durations and length

Row	Dataset	MHSA	T	PE	A Acc.	V Acc.	AV Acc.	A Entropy	V Entropy	A-V Entropy Diff
1	CREM-D	✓	6	✓	57.5	51.7	67.2	0.59	0.44	0.15
2		✗	6	✓	57.6	51.4	64.4	0.69	0.33	0.36
3		✓	12	✓	57.0	50.5	66.5	0.54	0.38	0.16
4		✗	12	✓	57.2	51.1	62.3	0.76	0.28	0.48
5		✓	6	✗	53.5	49.8	65.0	0.37	0.28	0.09
6		✗	6	✗	56.0	49.0	61.8	0.54	0.24	0.30
7		✓	12	✗	51.6	49.5	63.6	0.34	0.27	0.07
8		✗	12	✗	55.6	48.6	58.3	0.65	0.20	0.44
9	RAVDESS	✓	8	✓	59.2	58.2	76.3	0.41	0.32	0.09
10		✗	8	✓	61.6	55.3	70.6	0.73	0.25	0.47
11		✓	16	✓	58.5	55.8	74.4	0.40	0.30	0.10
12		✗	16	✓	59.0	56.2	68.8	0.90	0.25	0.65
13		✓	8	✗	55.4	54.4	75.2	0.27	0.25	0.02
14		✗	8	✗	60.7	56.0	69.4	0.60	0.21	0.39
15		✓	16	✗	54.7	54.2	72.4	0.26	0.23	0.02
16		✗	16	✗	57.2	54.7	65.2	0.77	0.20	0.57

It also shows unimodal and multimodal accuracies and the entropies of modalities? predictions. Here, MHSA represents the attention mechanisms within the proposed framework. Accuracies of emotion recognition through Audio-only, Video-only, and Audio-Video are referred to as A, V, and AV, respectively

76:3% and 67:2% for RAVDESS and CREMA-D, respectively. PE enhances the performance since it provides the topology with temporal information, where the improvement over using only the attention is at least 1%. This information is further utilized through the MHSA layer. Moreover, PE?s impact is more obvious when the number of time windows is large. For example, in the case of RAVDESS, with T = 16, PE improved the performance by 2:6%. In the baseline results of the framework, in case of not using both PE and attention, the performance drops by at least 5% ad 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%.

4.2.1 Impact of attention mechanisms on the bimodal recognition

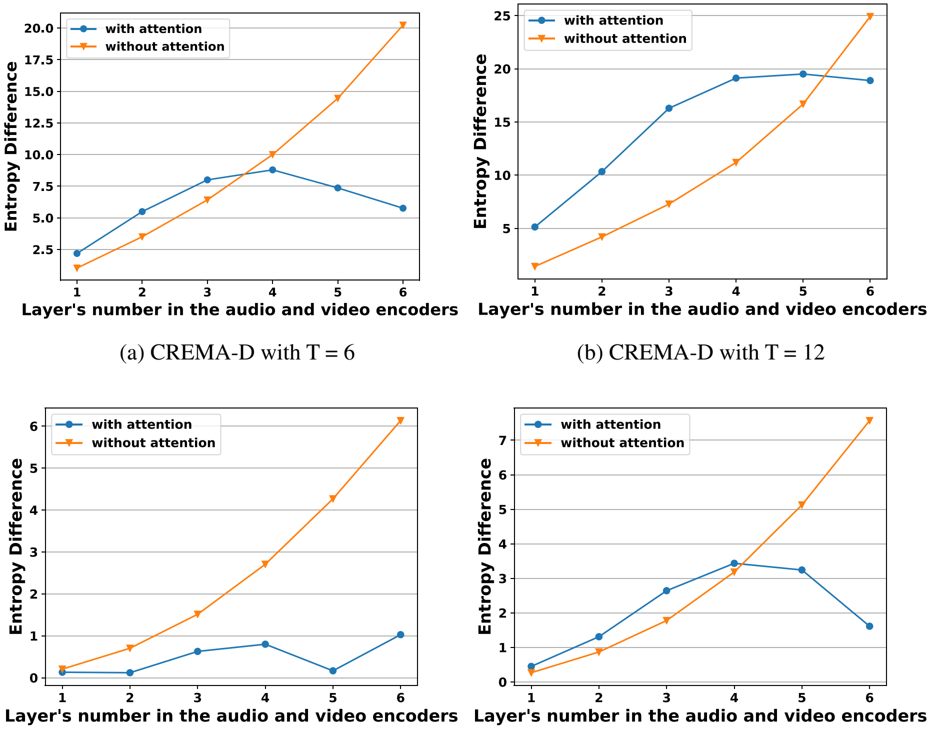
We observed that the differences in performance with and without attention mechanisms in the unimodal case are much lower than the corresponding figures in the multimodal case. For example, for CREMA-D, as shown in row 1 and column 6 of Table 1, using only the audio modality with PE and MHSA yields 57.5%. In the baseline case where PE and MHSA are not applied (shown in row 6), the recognition accuracy is 56.0%. However, in the audio-visual perception, which is shown in column 8, emotion recognition rates when employing both PE and MHSA and the baseline (i.e., when removing both PE and MHSA) are 67.2% (row 1) 61.8% (row 6), respectively. We can see that attention mechanisms with PE enhanced the multimodal performance significantly. For this reason, we want to study the framework's performance further to check the reason behind the variant improvements and how attention mechanisms helped the multimodal fusion over the unimodal perception.

To study how the bimodal fusion varies with and without attention mechanisms, we adopted the entropy to check the underlying agreement in audio and video scores. This is because entropy measures the average level of uncertainty (information) in variables' outcomes. In particular, we calculated the entropies of the audio and video predictions, since we applied late fusion on these predictions. Next, we measured the differences between the entropies of each of the audio and video samples.

Entropies of Layers Embeddings Furthermore, we measured the entropies of the embeddings at each layer of the framework (the encoder consists of $l = 6$ stacked blocks, as explained in Section 3.2). We first apply softmax normalization on the audio-visual representations, and then the average entropies ($H^{m\in\{audio(a),video(v)\}}$) of $d-$dimensional audio-visual representations across the layers were calculated as follows:

$$H^{(m)} = -\sum_{j=0}^{l}\sum_{i}^{d} \boldsymbol{h}_i^{(j)(m)} \log(\boldsymbol{h}_i^{(j)(m)}). \tag{6}$$

Figure 4 depicts a detailed comparison between the entropies of audio and video embeddings across the framework layers. The figure reveals that the usage of MHSA helped decrease the difference between audio and video entropies, while in the baseline, the difference is increasing in the later layers. The figure reveals that the usage of MHSA helped decrease the difference between audio and video entropies, while in the baseline, the difference is increasing in the later layers. To conclude, attention mechanisms helped the framework bring entropy measurements of audio and video embeddings and predictions closer to each other. Hence, these low differences between these entropies enhanced the bimodal certainty and improved the performance of audio-visual fusion.



(a) CREMA-D with T = 6

(b) CREMA-D with T = 12

(c) RAVDESS with T = 8

(d) RAVDESS with T = 16

Fig. 4 Entropy differences between audio and video embeddings in the case of the baseline and the framework with attention

4.2.2 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 (of MHSA), with 8, and 6 time windows for CREMA-D and RAVDESS, respectively. In CREMA-D, our approach outperformed the results in [9, 19]. Besides, in RAVDESS, the proposed topology resulted in higher accuracy than those in the literature but less than the recognition rate obtained through human perception. In [9], the performance (65.0% and 58.3.% accuracies, for CREMA-D and RAVDESS, respectively) was obtained by combining facial and audio temporal features with LSTM using Dual-Attention. In [19], a metric learning approach (Multimodal Emotion Recognition Metric Learning (MERML)) was applied to fuse audio-video modalities. MATER's results show its efficiency for enhanced joint multimodal learning and fusion. In [4], Athanasiadis et al. employed transfer learning between audio and visual modalities using Generative Adversarial Networks (GANs), which is a more challenging learning task than the one addressed in this study. In [18], the proposed procedure applied LSTM-based end-to-end Deep Metric Learning (DML) on raw audio and visual data, using SoundNet [6] for audio mappings and I3D [11] for visual mappings. These paradigms achieved accuracies of 74.3% and 70.1% on CREMA-D and RAVDESS, respectively. Besides, when replacing LSTM by the Transformer's encoders and attention

Table 2 Audio-Video average accuracies (%) of MATER and other related work

Approach	CREMA-D	RAVDESS
Human Perception: AV	74.8	80.0
Dual Attention with LSTM: AV [9]	65.0	58.3
MERML [19]	66.5	66.3
Visual (I3D [11])+ Audio (SoundNet [6]) + LSTM + triplet loss [18]	74.3	70.1
Visual (I3D [11]) + Audio (SoundNet [6]) representations [18]) + attention + softmax loss	74.1	74.6
Supervised Generative Adversarial Networks [4]	52.2	47.11
Temporal aggregation using CNNs [32]	84.0	78.7
MATER: $AV_{+PE+MHSA}$	67.2	76.3

mechanisms, and the triplet loss of the DML by softmax loss function (in particular, we use cross-entropy), we obtained high performance of 74.1% and 74.6% accuracies on CREMA-D and RAVDESS, respectively. In [32], Radio et al. used temporal aggregation through CNNs, and utilized data augmentation to enrich the pool of training data. Nonetheless, we notice that replacing the audio-visual mappings of SoundNet [6] and I3D [11], with the extracted embeddings in this study has the following advantages:

- In this paper, we employed smaller time windows in comparison to previous studies. Also, SoundNet and I3D work best with larger time windows in the end-to-end learning paradigm. Nonetheless, the *interpretability* of the attention mechanisms on top of these deep and large models is not feasible. As a result, we opt for replacing the audio and visual mappings with the extracted embeddings as elaborated in Section 3.1.
- MATER on the pre-extracted embeddings of small time windows offers *explainability* of attention mechanisms for fusing audio-visual cues and spotting the important time segments. Consequently, using MATER, it is attainable to analyze the obtained results extensively, as demonstrated in the evaluations of this and the subsequent sections.

5 Handling noisy data with attention mechanisms

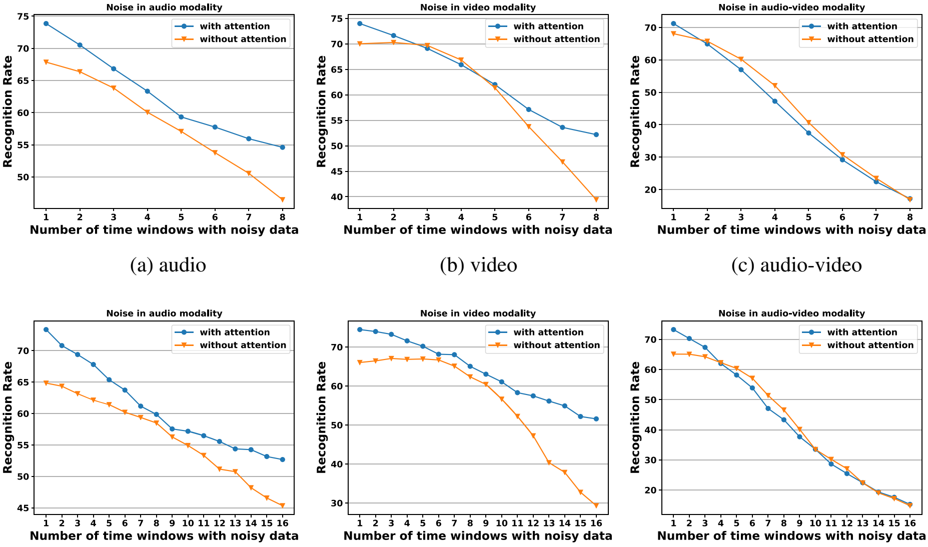
The evaluations in this section aim to investigate the framework's robustness when trained and evaluated with noisy data. The study checks how attention mechanisms help the model under challenging conditions. In the following subsections, we present two types of evaluations. The first set of experiments examines the models that were trained with *"noise-free"* but are evaluated with noisy embeddings in some or all time windows. The second set of experiments introduces re-training and evaluating the MATER framework with noisy data. In the results of this section and the following section (Section 6), we use two types of models "with attention" and "without attention". The models "without attention" refer to the baseline as described in Section 4.2, where PE and MHSA are removed. Models with attention refer to the framework, where PE and MHSA are kept.

5.1 Noise injection into noise free models

In this section's evaluations, the noise was injected at different numbers of time windows in the three settings, i.e. audio only, video only, and audio-video fusion. For instance, a number of 1 and up to T time windows were replaced with noise in audio-only, video-only, and on both audio-video fusion. The evaluated models were trained with "noise-free" data. Since noise is a random signal and was injected in random places in the overall data, we opted for repeating the experiment several times (here, 10) to establish reliable figures regarding performance. Hence, the following reported results are the average results of the 10 evaluations. The noise injection refers to replacing time windows' embeddings with random signals sampled from Gaussian.

5.1.1 Evaluation results

Figures 5 and 6 demonstrate the multimodal recognition rates of the aforementioned scenario. We notice that when both modalities are employed in the case of using the attention mechanisms, the framework's performance is degraded with noisy data quicker than the baseline, as displayed in Figs. 5c, 5f, 6c, and 6f. This is an expected outcome, since in the case of using MHSA, time windows are interacting with each other (as explained in Section 3.2), while in the baseline, the multimodal prediction mainly relies on the time windows' aggregated predictions. Besides, an interesting observation of this evaluation is that when injecting noise in one modality, the framework could still depend on the other



(a) audio

(b) video

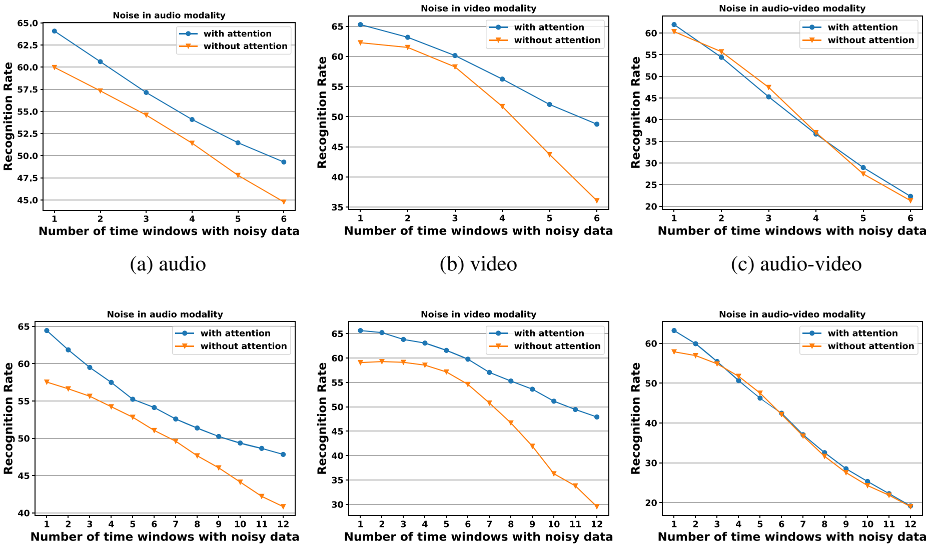
(c) audio-video

(d) audio

(e) video

(f) audio-video

Fig. 5 RAVDESS: The results of evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The adopted models in these figures were trained with both audio and video modalities jointly. Nonetheless, in the evaluation phase, as indicated by each subfigures title, the noise was placed in audio-only, video-only, or both modalities' embeddings



(a) audio (b) video (c) audio-video

(d) audio (e) video (f) audio-video

Fig. 6 CREMA-D: The results of evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The adopted models in these figures were trained with both audio and video modalities jointly. Nonetheless, in the evaluation phase, as indicated by each subfigures title, the noise was placed in audio-only, video-only, or both modalities' embeddings

modality without losing its unimodal performance. For instance, when placing noise in all the time windows of video modality (as shown in Figs. 5b, 5e, 6b, and 6e), the multimodal performance of the framework was similar to the reported audio results in Table 1.

Algorithm 1 Noise Injection Algorithm.

```
procedure RETRAINING MATER WITH NOISY DATA($\mathbb{D}$)
Inputs:
    Formulate the method as shown in Figure 1 and described in Section 3
    Audio ($f^{(a)}(x^{(a)})$) and visual ($f^{(v)}(x^{(v)})$) embeddings
Initialization:
    Number of time windows (T)
    Noisy modalities: M ∈ {audio, video, both audio-visual}
    Training and evaluation parameters as described in Section 4.1
Training:
    for m = 1:M do
        Re-train MATER (with attention and the baseline models) using noise in $m$
        Use the same parameters and details as described in Section 4.1
        In each iteration during the training, pick a random number: $T_{noisy} \in \{1, ..., \frac{T}{2}\}$,
and replace the resulted $T_{noisy}$ time  windows with noise
    end for
Evaluation:
    Evaluate the obtained baseline and attention models using noise in $m$ modality (with
the same settings)
end procedure
```

5.2 Retraining the framework with noisy time windows

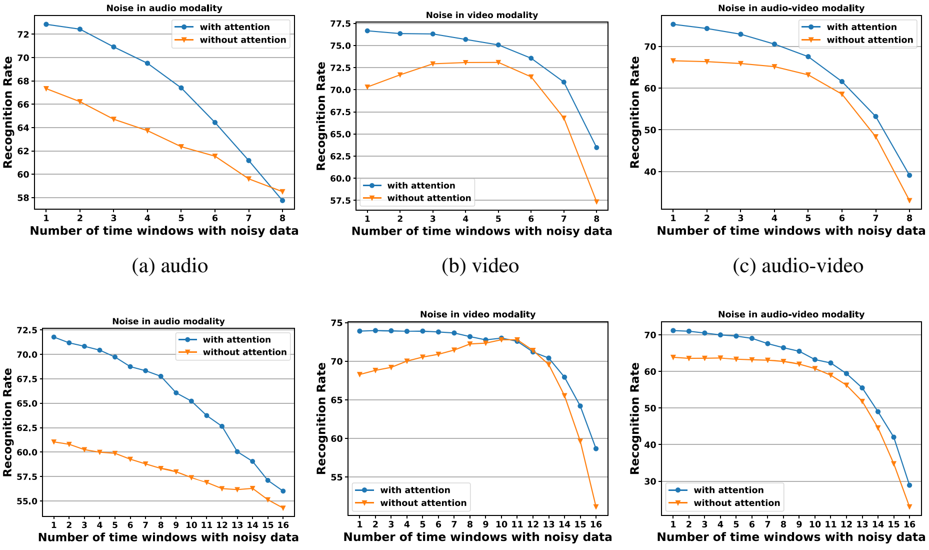
The motivation behind the retraining of the framework with noise is to check the model's performance when it is exposed to similar conditions during the training and testing. This study allows us to monitor the validity and the robustness of MATER's performance when using attention mechanisms, under challenging settings.

5.2.1 Retraining scenarios

The framework was retrained (from the scratch) using noise in audio-only, video-only, and both audio-video modalities, separately. In other words, we obtain three models from the proposed method, each model has noisy in audio, video, or audio-video embeddings. We injected the noise under the following restriction: the maximum number of noisy time windows is set to half the total number of the time windows. For instance, if T (time windows' number) is 8, the number of the noisy time windows might range between 1 and 4, and the specific number is chosen randomly at each iteration during the training process. Moreover, their positions are scattered randomly across the time windows. These restrictions were placed to allow the framework to converge during the training process, despite employing noisy embeddings. Algorithm 1 summarizes the noise injection scenarios in the training processes.

5.2.2 Evaluation scenarios

Each baseline and the framework with the attention mechanisms have three models trained with noise for each modality. In the evaluation process, if a model was trained with noise,



(a) audio (b) video (c) audio-video

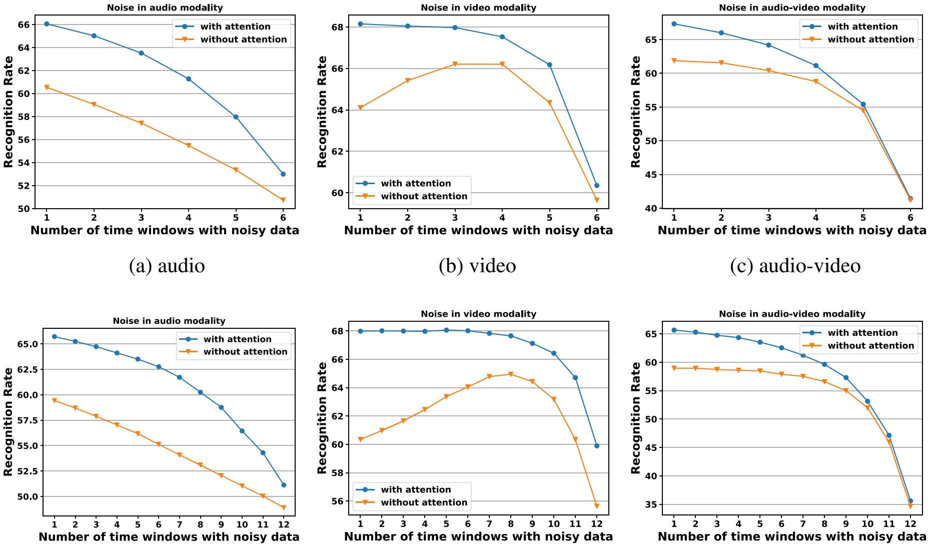
(d) audio (e) video (f) audio-video

Fig. 7 RAVDESS: The results of retraining and evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The title of each sub-figure indicates the modality in which noise was used

e.g. in audio, during the testing process, the noise was used only in the same modality. In terms of the number of noisy time windows for noise injection and their positions, the noise was injected in 1 to T time windows as shown in Figs. 7 and 8. Moreover, as the used embeddings contain noise, the models' evaluations were performed 10 times and the reported results are the average results of these evaluations. Figures 7 and 8 illustrate the results of the retraining and testing schemes. Without exception, we notice that when the framework is trained with noisy data and tested similarly, its performance is significantly robust, especially when using the attention mechanisms. This is in contrast to the case of evaluating the models which were obtained using *"noise-free"* data during training. This consistency implies the adaptability of the method to more challenging settings. In addition, it demonstrates that the attention mechanisms can handle the instances of noisy time windows due to the re-training procedure. In addition, interestingly, Figs. 7b, 7e, 8b, and 8e show how the video modality can recover from noise, demonstrating that there is not a specific range or number of time windows that are more important than others. A further interpretation could be drawn that the attention mechanisms played a bigger role in audio modality than the one in video modality, as the audio modality's strength is concentrated in the middle of video clips (as will be shown in the evaluations of the following section, Section 6).

6 Evaluations of emotion recognition overtime

This section aims to study the impact of temporal information on emotion recognition accuracy and the impact of attention mechanisms, mostly from a qualitative point of view, instead of merely performance-driven experiments.



(a) audio (b) video (c) audio-video

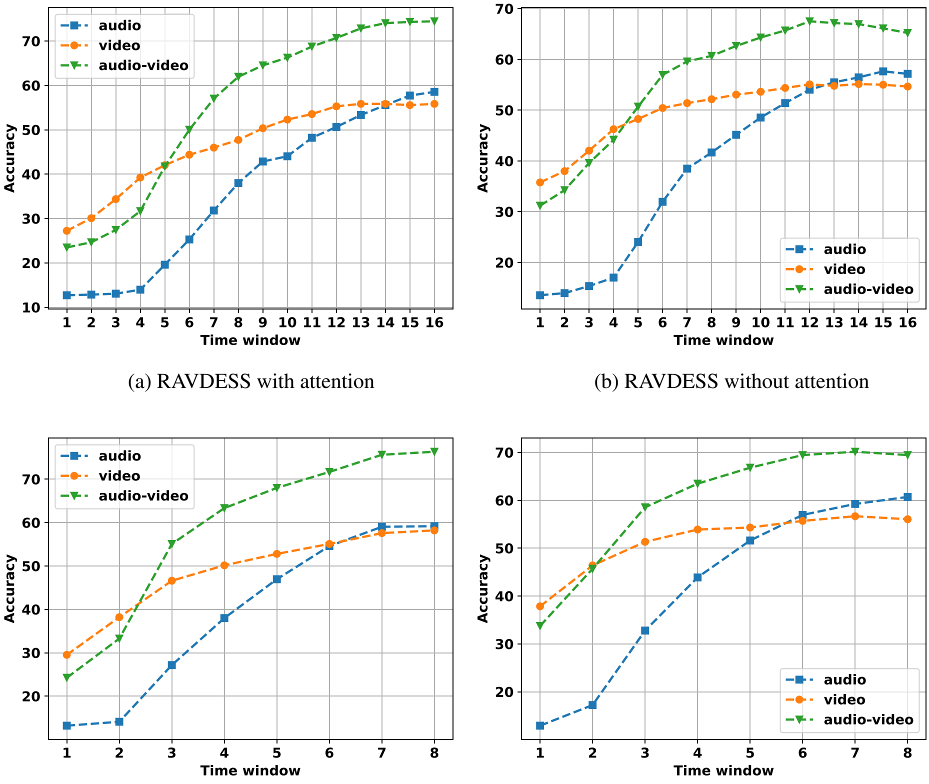
(d) audio (e) video (f) audio-video

Fig. 8 CREMA-D: The results of retraining and evaluating the framework with noisy data at different numbers of time windows. The x-axis shows the number of time windows which were replaced by Gaussian noise. The title of each sub-figure indicates the modality in which noise was used

6.1 Incremental emotion perception

Humans recognize emotions at different rates across modalities [10]. For example, the developers of CREMA-D [10] studied human raters' recognition speed of emotions. They found that the raters need more time to recognize emotions through vocal expressions than the time needed to recognize emotions through facial expressions. Motivated by these findings, we study the differences between modalities response times in the MATER framework.

First, we examine how the proposed framework captures the temporal display of emotions through audio and video modalities. For instance, Fig. 9 shows the results of the framework up until each time window (t). Specifically, in Fig. 9a, the performance at time window 3 implies that the audio, video, or audio-video embeddings of the first, second, and third time windows were used in the framework. We notice that the performance of MATER with PE and MHSA peaks at the last time window. It means that the framework could make use of the embeddings from all the time windows. However, this evaluation shows that, in



(a) RAVDESS with attention

(b) RAVDESS without attention

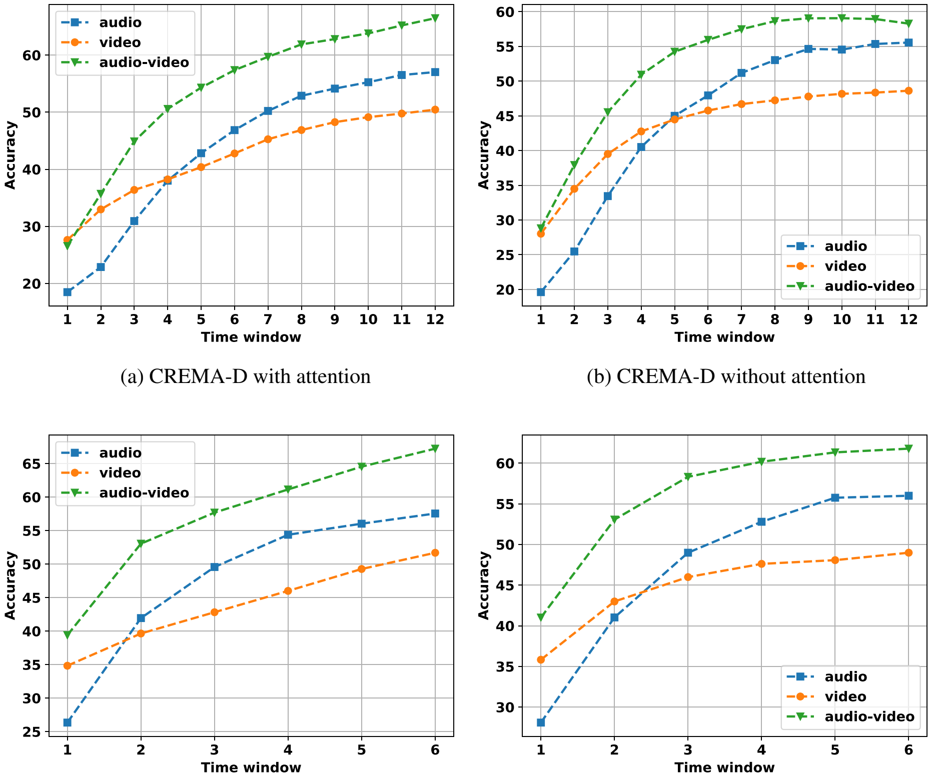
(c) RAVDESS with attention

(d) RAVDESS without attention

Fig. 9 RAVDESS: incremental performance results. These results show the incremental presentations of the embeddings to the framework, for audio-only, video-only, and their multimodal fusion. For example, when $T = 3$, it means that the corresponding accuracy shows the results when the first three time windows were used

the baseline case, where PE and MHSA are not used, the performance drops prior to the last time windows. The final time windows could be useful and the attention mechanisms are able to capture their importance for emotion recognition. These findings are observed in both datasets, using time windows with varying lengths and numbers. Furthermore, the presented results reveal that, in most of the cases, the video modality reached a plateau state earlier than the audio modality.

In addition, Figs. 9 and 10 show that the recognition rates of audio modality in the initial time windows are lower than the video modality ones. Specifically, the audio modality requires more time to achieve comparable performance (or even higher in some cases) with the video modality. Also, the recognition rates of emotions through video modality usually rise sooner than the audio modality ones. Moreover, as mentioned previously, there is usually a drop in the network's performance in case of not using the attention. However, the attention mechanisms efficiently bridge the gap between audio and video modalities' performances through the time windows. Subsequently, the reduced gap contributed to accurate bimodal emotion recognition.



(a) CREMA-D with attention

(b) CREMA-D without attention

(c) CREMA-D with attention

(d) CREMA-D without attention

Fig. 10 CREMA-D: incremental performance results. These results show the incremental presentations of the embeddings to the framework, for audio-only, video-only, and their multimodal fusion. For example, when $T = 3$, it means that the corresponding accuracy shows the results when the first three time windows were used

6.2 Modalities response time

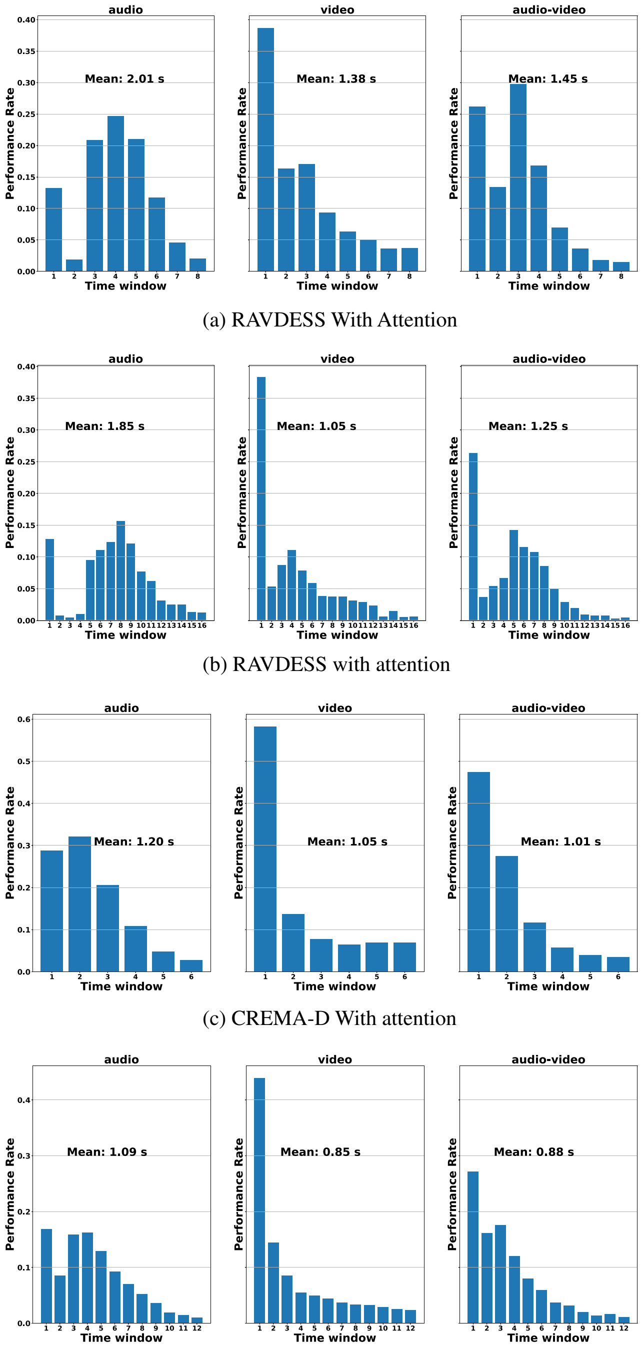
People's response time could vary according to the presented modalities [8, 10, 25]. The response time depends on the emotion's intensity and duration. The response time refers to the minimum duration required for emotion perception and recognition. For example, the authors in CREMA-D [10] studied the mean response time of human perception of emotions through audio only, video only, and their combination. The study reported that, on average, the mean response times of human raters for emotion recognition through audio only, video only, and the bimodal audio and video perception are 2.98, 2.05, and 1.95 seconds, respectively. The results show that the recognition speed through audio modality is at least one second lower than video only or both audio-visual cues. Furthermore, in RAVDESS [27], human raters' average response times are 1.55, 1.31, and 1.32 seconds, for audio only, video only, and the audio-video bimodal perception, respectively.

In human rating, the response time can be measured by taking the time when the raters first identify an emotion of a video clip. However, in automatic emotion recognition, due to technical requirements, the response time could be measured with the smallest possible time window of a framework. In our case, this duration is either 0.5 or 0.25 seconds. In this paper, we only consider the correctly classified samples to measure modalities' response time. As a reference measurement, we use the duration of a time window where the video content was correctly predicted for the first time. For example, if a video was correctly classified through the audio modality in the 4^{th} time window of 0.25 second duration, then the response time is set to 1 second. The sample is then added to a histogram that demonstrates the ratio of correctly classified videos in this time window. Figure 11 shows performance rates of different time windows, related to their contributions (ratios) in recognizing emotions as a function of time. E.g. in Fig. 11a (audio-video), emotions were identified correctly already since time window 1 in 26% of the cases, while they were correctly classified starting from the second time window in a 14% of the time.

According to Fig. 11, the video modality relies on the initial time windows. However, the middle time windows are important in the audio modality. Moreover, the figure displays the mean response time per-modality, for RAVDESS and CREMA-D. For example, in RAVDESS, Fig. 11a shows that the mean response times are 2.01, 1.38, and 1.45 seconds, for audio-only, video-only, and the audio-video fusion, respectively. This implies the fact that audio modality requires more time than video modality for emotion recognition. These findings are consistent among different figures. Moreover, they are similar to those obtained through the response time of human perception of emotion. In the automatic bimodal audio-video emotion recognition, we observed that the delay in the audio mean response time slightly impacts the bimodal one, making the response time of the video modality the shortest one among the three settings.

7 Conclusions

The research in this paper studies the importance of exploiting audio-video signals' temporal strength for emotion recognition. We utilize the attention mechanisms 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 transformer attention mechanisms significantly improves the performance over the baseline (the same topology without the attention mechanism's selective character). Our results demonstrate the importance of weighing the contribution of each modality separately and of different time



(a) RAVDESS With Attention

(b) RAVDESS with attention

(c) CREMA-D With attention

(d) CREMA-D with attention

Fig. 11 RAVDESS: incremental performance results. These results show the incremental presentations of the embeddings to the framework, for audio-only, video-only, and their multimodal fusion. For example, when T = 3, it means that the corresponding accuracy shows the results when the first three time windows were used

windows. Besides, the framework gave more insights concerning the multimodal interaction and presentation of audio-visual cues. For instance, the examination of the framework showed that the attention mechanism helps in capturing the temporal presentation of the audio-video embeddings, resulting in the best performance in the last time window. In addition, the proposed framework provides a close interpretation and insights regarding multimodal emotion recognition, making use of visual and audio cues. Specifically, the presented comprehensive evaluations led to the following conclusions:

- The joint modelling of audio-visual cues using the attention mechanisms contributes to bringing the entropies of the audio and video modalities closer. This modelling increased the certainty in the multimodal predictions, which enhanced the multimodal perception compared to the baseline model.
- Attention mechanisms improve the robustness of the proposed framework. Evaluating the framework with noise demonstrated that the method is robust when exposed to similar conditions during the training and testing procedures. Finally, injecting noise into the framework showed that the audio modality is more vulnerable to noisy data, while the video modality is more robust against noise.
- Attention mechanisms utilize embeddings from overall time windows and capture how video and audio modalities behave across time. For example, the examination of the framework showed that the attention mechanisms help capture the incremental presentation of the audio-video embeddings, resulting in the best performance in the last time window.

The future direction of the work includes employing attention mechanisms for more challenging datasets, such as those recorded or gathered in the wild. Future work can also employ attention mechanisms within computational frameworks when building feature representations, to benefit from these techniques in obtaining robust representations, taking into consideration informative data points spatially and temporally.

Declarations The authors have no relevant financial or non-financial interests to disclose.

References

1. Abu-El-Haija S, Kothari N, Lee J, Natsev A, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: a large-scale video classification benchmark. arXiv:1609.08675
2. Afouras T et al (2018) Deep audio-visual speech recognition. IEEE Trans Pattern Anal Mach Intell
3. Albanie S, Vedaldi A (2016) Learning grimaces by watching tv BMVC
4. Athanasiadis C, Hortal E, Asteriadis S (2020) Audio–visual domain adaptation using conditional semi-supervised generative adversarial networks. Neurocomputing 397:331–344
5. Aubergé V, Cathiard M (2003) Can we hear the prosody of smile? Speech Commun 40(1-2):87–97. https://doi.org/10.1016/S0167-6393(02)00077-8
6. Aytar Y, Vondrick C, Torralba A (2016) Soundnet: learning sound representations from unlabeled video. In: Advances in neural information processing systems, pp 892–900

7. Baltrušaitis T, Ahuja C, Morency L-P (2019) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443
8. Barkhuysen P, Krahmer E, Swerts M (2010) Crossmodal and incremental perception of audiovisual cues to emotional speech. Lang Speech 53(1):3–30
9. Beard R, Das R, Ng RW, Gopalakrishnan PK, Eerens L, Swietojanski P, Miksik O (2018) Multimodal sequence fusion via recursive attention for emotion recognition. In: Proc of the 22nd conf on computational natural language learning, pp 251–259
10. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R (2014) Crema-d: crowd-sourced emotional multimodal actors dataset. IEEE Trans Affective Comput 5(4):377–390. https://doi.org/10.1109/TAFFC.2014.2336244. arXiv:NIHMS150003
11. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
12. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. arXiv:1405.3531
13. Cosentino S, Randria EI, Lin J-Y, Pellegrini T, Sessa S, Takanishi A (2018) Group emotion recognition strategies for entertainment robots. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 813–818
14. D'Mello S, Kappas A, Gratch J (2018) The affective computing approach to affect measurement. Emotion Rev 10(2):174–183. https://doi.org/10.1177/1754073917696583
15. D'mello SK, Kory J (2015) A review and meta-analysis of multimodal affect detection systems. ACM Comput Surveys 47(3):1–36. https://doi.org/10.1145/2682899
16. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, vol 1 (long and short papers), pp 4171–4186. Association for computational linguistics. https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
17. Ekman P, Friesen WV, O'sullivan M, Chan A, Diacoyanni-Tarlatzis I, Heider K, Krause R, LeCompte WA, Pitcairn T, Ricci-bitti PE et al (1987) Universals and cultural differences in the judgments of facial expressions of emotion. J Personality Social Psychology 53(4):712
18. Ghaleb E, Popa M, Asteriadis S (2019) Multimodal and temporal perception of audio-visual cues for emotion recognition. In: 2019 8th International conference on affective computing and intelligent interaction (ACII). IEEE, pp 552–558
19. Ghaleb E, Popa M, Asteriadis S (2019) Metric learning based multimodal audio-visual emotion recognition. IEEE MultiMed
20. Goodfellow IJ, Erhan D, Carrier PL et al (2013) Challenges in representation learning: a report on three machine learning contests. In: Int conf on neural information processing. Springer, pp 117–124
21. Hershey S et al (2017) Cnn architectures for large-scale audio classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 131–135
22. Hori C et al (2019) 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, pp 2352–2356
23. Kappas A (2013) Social regulation of emotion: messy layers. Front Psychol 4:51
24. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: Proc of the IEEE Conf on computer vision and pattern recognition (CVPR). IEEE, pp 1867–1874
25. Kim Y, Provost EM (2016) Emotion spotting: discovering regions of evidence in audio-visual emotion expressions. In: Proceedings of the 18th ACM international conference on multimodal interaction. ACM, pp 92–99
26. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations (ICLR)
27. Livingstone SR, Russo FA (2018) 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 13(5):0196391
28. Mano LY, Faiçal BS, Nakamura LH et al (2016) Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition. Comput Commun 89:178–190
29. Ngiam J, Khosla A, Kim M, Nam J et al (2011) Multimodal deep learning. In: Proc of the 28th int Conf on machine learning (ICML-11), pp 689–696
30. Picard RW (2000) Affective computing. MIT Press
31. Plutchik R (2001) The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
32. Radoi A, Birhala A, Ristea N-C, Dutu L-C (2021) An end-to-end emotion recognition framework based on temporal aggregation of multimodal information. IEEE Access 9:135559–135570

33. Ringeval F, Schuller B, Valstar M et al (2019) AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: AVEC 2019 - Proceedings of the 9th international audio/visual emotion challenge and workshop, co-located with MM 2019 (Avec), pp 3–12. arXiv:1907.11510. https://doi.org/10.1145/3347320.3357688
34. Rouast PV, Adam M, Chiong R (2019) Deep learning for human affect recognition: Insights and new developments. IEEE Trans Affective Comput
35. Shi Y, Siddharth N, Paige B, Torr P (2019) Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: Advances in neural information processing systems, pp 15718–15729
36. Vaswani A et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
37. Wang W, Tran D, Feiszli M (2020) What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12695–12705
38. Wu C-H, Huang Y-M, Hwang J-P (2016) Review of affective computing in education/learning: trends and challenges. Br J Educ Technol 47(6):1304–1323
39. Wu Z, Zhang X, Zhi-Xuan T, Zaki J, Ong DC (2019) Attending to emotional narratives. In: 2019 8th International conference on affective computing and intelligent interaction (ACII). IEEE, pp 648–654
40. Xu J, Yao T, Zhang Y, Mei T (2017) Learning multimodal attention lstm networks for video captioning. In: Proceedings of the 25th ACM international conference on multimedia, pp 537–545
41. Xu H, Zhang H, Han K, Wang Y, Peng Y, Li X (2019) Learning alignment for multimodal emotion recognition from speech. In: INTERSPEECH, pp 3569–3573. https://doi.org/10.21437/Interspeech.2019-3247

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Esam Ghaleb[1] · **Jan Niehues**[1] · **Stylianos Asteriadis**[1]

Jan Niehues
jan.niehues@maastrichtuniversity.nl

Stylianos Asteriadis
stelios.asteriadis@maastrichtuniversity.nl

[1] Department of Data Science and Knowledge Engineering, Maastricht University,
Maastricht, Netherlands