Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning

Bogdan Mocanu, Ruxandra Tapu, Titus Zaharia

Contents lists available at ScienceDirect

Image and Visual Computing

Multimodal Emotion Recognition using Cross Modal Audio-Video Fusion with Attention and Deep Metric Learning

Bogdan Mocanu[a], Ruxandra Tapu[b]*, Titus Zaharia[b]

[a]*Department of Telecommunications, Faculty of ETTI, University "Politehnica" of Bucharest, Romania*
[b]*Institut Polytechnique de Paris, Télécom SudParis, ARTEMIS Department, 9 rue Charles Fourier, 91000 Évry, France*

Abstract

In the last few years, the multi-modal emotion recognition has become an important research issue in the affective computing community due to its wide range of applications that include mental disease diagnosis, human behavior understanding, human-machine/robot interaction or autonomous driving systems. In this paper, we introduce a novel end-to-end multimodal emotion recognition methodology, based on audio and visual fusion designed to leverage the mutually complementary nature of features while maintaining the modality-specific information. The proposed method integrates spatial, channel and temporal attention mechanisms into a visual 3D convolutional neural network (3D-CNN) and temporal attention into an audio 2D convolutional neural network (2D-CNN) to capture the intra-modal features characteristics. Further, the inter-modal information is captured with the help of an audio-video (A-V) cross-attention fusion technique that effectively identifies salient relationships across the two modalities. Finally, by considering the semantic relations between the emotion categories, we design a novel classification loss based on an emotional metric constraint that guides the attention generation mechanisms. We demonstrate that by exploiting the relations between the emotion categories our method yields more discriminative embeddings, with more compact intra-class representations and increased inter-class separability. The experimental evaluation carried out on the RAVDESS (*The Ryerson Audio-Visual Database of Emotional Speech and Song*), and CREMA-D (*Crowd-sourced Emotional Multimodal Actors Dataset*) datasets validates the proposed methodology, which leads to average accuracy scores of 89.25% and 84.57%, respectively. In addition, when compared to state-of-the-art techniques, the proposed solution shows superior performances, with gains in accuracy ranging in the [1.72%, 11.25%] interval.

Keywords: Spatial attention; Channel attention; Temporal attention; Cross-modal fusion; Emotional metric constraint

* Corresponding author. Tel.: +33 1 60 76 43 65.
E-mail address: ruxandra.tapu@telecom-sudparis.eu

Please cite this article as: First author et al., Article title, Image and Visual Computing (2022)

1. Introduction

Emotions are omnipresent in any moment of our daily life and play an important role in human interaction. They reveal intentions, carry empathy and, in some cases, allow us to transmit information more effectively. The nature of human communication is inherently multi-modal and includes both linguistic and paralinguistic components. The former refers to the spontaneous conversational exchange (verbal content), while the latter refers to characteristics that are performed either intentionally or subconsciously that involve aspects related to facial expressions, body gestures, or vocal features such as speaking rate or intonation.

In the last decades, the issue of emotion analysis has drawn an increasing interest in various research fields that involve artificial intelligence, psychology, neurosciences, medicine and autonomous driving systems. Within this framework, a large variety of both continuous and discrete emotional models [1] can be considered. The dimensional models represent emotions as a continuous spectrum (*i.e.*, activation and valence), while the latter one quantifies emotions into a set of discrete categories. In the pioneering work of Eckman and Friesen [2], [3], the discrete emotions are qualified into a set of six basic states (*i.e.*, fear, disgust, sadness, happiness, anger, and surprise) that human can perceive similarly regardless of their regional, ethnical, or cultural differences.

Automatic discrete emotion identification in real world scenarios is a very difficult and challenging task, which requires a high level, cognitive understanding of both verbal and non-verbal communication. In addition, the emotion recognition process is highly subjective because people can interpret emotion differently, depending on the environmental factors or current state of mind.

Human emotions can be identified through various modalities, including facial expression, speech data, body gestures, text, or physiological data (*e.g.*, electrocardiogram, electroencephalogram...), which typically carry complementary information. Although various studies have employed more complex modalities, audio and visual features are still the primary contact-free modalities used to convey emotions. Within the context of autonomous driving systems (ADS), it is becoming increasingly important to recognize continuously the driver's emotions/state of mind, allowing thus the intelligent vehicle to respond in an optimized manner to the user needs and consequently select the optimal driving mode.

In this paper, we introduce a novel multimodal emotion recognition framework based on A-V information designed to leverage the mutually complementary nature of features while maintaining the modality-specific information. The proposed system can be easily integrated in any ADS, being able to recognize the driver's emotional state among a well-defined set of discrete emotional categories. We notably exploit the two non-invasive sources of information, which are the visual and audio signals. In real life scenarios, such data can be straightforwardly acquired with the help of a video camera installed inside the vehicle.

The main contributions of the paper are the following:

(1). A deep learning-based multimodal emotion recognition framework that includes various self-attention mechanisms. The system performs an independent analysis over the audio and video channels to extract discriminative inter-modal characteristics. For the visual channel, three different types of attention methods (including spatial, channel-wise and temporal) are employed, while for the audio channel solely the temporal attention is used.

(2). A novel model-based fusion strategy which uses cross-attention and determines the interaction between A-V representations, while capturing the intra-modal correlations between modalities.

(3). A learnable emotional metric that extends the traditional triplet loss function with an additional constraint, which enables to generate polarity-preserved attention maps. By taking into consideration the relations between the emotion categories more discriminative embeddings are obtained, with more compact intra-class representations and increased inter-class separability.

(4). An extensive objective evaluation, carried out on the RAVDESS [4] and CREMA-D [5] datasets. The experimental evaluation demonstrates that the proposed methodology achieves better performances when compared with salient state-of-the-art methods.

The paper is organized as follows. Section 2 presents the state-of-the-art dedicated to discrete emotion recognition, emphasizing the main families of existing approaches with related strengths and limitations. Section 3 describes the proposed methodology and details the key steps involved. The experimental setup, with training protocol, datasets and experimental results obtained are presented and discussed in Section 4. Finally, Section 5 concludes the paper and opens some perspectives of further work.

2. Related work

In this section, we focus our attention solely on discriminative algorithms that have been introduced in the last couple of years. Depending on the type of modalities used to convey emotions, we can identify speech-based, visual-based and multi-modal emotion recognition frameworks.

2.1. Speech-based emotion recognition systems

The performance of speech-based emotion recognition (SER) systems is highly dependent on the low-level features representation extracted from the audio signal. Traditional approaches rely on hand-crafted audio features (*i.e.*, formant, loudness, linear productivity code...), speech statistics (*i.e.*, mean, median, standard deviation…) or specific descriptors (*i.e.*, Mel Frequency Cepstral Coefficient – MFCC). In order to automate the feature extraction process, dedicated frameworks such as OpenSmile [6] or Praat [7] have been introduced.

By using the MFCC with spectral centroids as input to a Support Vector Machine (SVM) classifier, the Bhavan *et al.* approach [8] achieves a 72.91% accuracy rate on the RAVEDESS [4] corpus. However, the main limitation of such global level acoustic feature representation is the reduced capacity to capture the speech variation dynamics along the complete length of the audio segment.

With the development of deep learning methods, SER models can extract relevant features from the audio stream without any human intervention. In [9], authors propose a system that infers affect-related salient features using convolutional neural networks (CNN). Pepino *et al.* [10] combines hand-crafted features and deep learning models (eGeMAPS) to represent the speech signal. For better accuracy, a transfer learning paradigm is employed, with the involved CNN pre-trained on different, large scale audio datasets. Both systems return superior accuracy scores when compared to traditional approaches.

To generate more precise emotions bi-classification results, in [11] discriminative feature spaces are constructed for two different emotions pairs. Based on the observation that some archetypical emotions are closer in the feature space representation a Naïve Bayes decision fusion classifier is proposed. In [12], a Siamese CNN architecture is used to increase the intra-class compactness and inter-class separability. Issa *et al.* [13] propose a 1D-CNN architecture fed with traditional descriptors such as MFCC, chroma-gram, mel-scale spectrogram, Tonnetz representations and spectral contrast to identify emotion from raw audio signals. Recently, in [14] and [15] the authors propose various emotion recognition frameworks based on CNN architectures extended with NetVLAD [14] and GhostVLAD [15] layers for feature aggregation.

In [16], recurrent neural networks (RNN) are proposed to learn short-time, frame-level acoustics as well as the appropriate temporal aggregation of features. Similarly, Tzinis *et al.* [17] introduced a RNN fed with both local and statistical (global) features. The system returns the best results when performing the analysis at different time scales and when extracting statistics at the level of the entire utterance. The triplet loss framework based on LSTM (Long Short-Term Memory) proposed in [18] is designed to learn a discriminative mapping of the feature embeddings. Based on the loss function, the intra-class distances can be reduced, while increasing the inter-class separation. In [19-21] various CNN models combined with LSTM [19] or with self-attention mechanisms [20-21] are used to increase the system robustness to noise or various compression artifacts.

Recently, the authors in [22] studied the confidence estimation of deep neural networks (DNN) within the context of SER having as goal to generate a model that outputs predictions only when it is sufficient confident. The method is based on a novel loss function, so-called confidence metric, computed between two types of emotion representation. Similarly, the sentiment-aware emotion recognition method introduced in [23] combines speech analysis with text-based sentiment classification. Other approaches [24-26] address the problem of DNN sensitivity to attacks of gradient distortions. In [24], Saurabh *et al.* are one of the first to propose the integration of a regularization term, derived from adversarial training, in order to smooth the model prediction. A different defense mechanism consists in training the model by augmenting the training dataset with attack samples, along with a feature similarity loss [25]. However, both methods [24], [25] can only guarantee protection against a specific (seen) type of attack at a pre-determined intensity. In [26], a self-supervised augmentation defense mechanism is proposed that learns to neutralize the gradient distortion without knowing the attack type, while in [27] the authors propose a model designed to generalize to unseen data with varying characteristics. Four different attributes have been considered to evaluate the performance of a CNN-LSTM

model: the corpus structure, the gender, the speakers, and the language. Finally, in [28] the magnitude and phase representations are used in the encoder part of a modified UNET architecture to cope with speech inputs acquired on the wild.

From the analysis of the state-of-the art, the following conclusions can be highlighted: (1). Existent approaches are still far from satisfactory in recognizing emotions from real audio streams. Most methods heavily depend on the training datasets that contain elicited or acted speech segments. (2). The systems are highly sensitive to the length of the speech utterance. Some authors propose constraining the audio signal to a fixed length representation by clipping or padding the utterance. However, such an approach reduces the discriminative power of the overall descriptor.

Let us now analyze the vision-based emotion recognition solutions.

2.2. Vision based emotion recognition systems

As in the case of SER systems, various visual descriptors can be extracted from the facial morphology to detect expressions in video streams. Nguyen *et al.* [22] extract the 68 facial landmark points to encapsulate meaningful visual information in order to discriminate between emotion classes. By using the facial landmarks, the authors construct 32 geometrical descriptors that are used to train a SVM classifier. In [23], a framework based on action units (AUs) is proposed. The model continuously detects the affective states using AUs that can be interpreted as an evolved version of landmark points. The AUs reflect the facial movement in time and not just the location for some regions of interest. The method is based on a stacked autoencoder network designed to predict discrete emotions.

With the recent advances in deep learning techniques, the vision-based emotion recognition systems using 2D/3D-CNN architectures that are receiving as input video frames/sequences, have returned higher recognition rates compared to traditional methods based on frame aggregation.

EmotionalDAN [24] is an example of 2D-CNN designed to solve the emotion, valence, and landmark recognition problem in one stage. Facial landmarks are here incorporated as a part of the classification loss function and an alignment-dedicated deep network is extended with a term related to facial features. The spatial transformers networks (STN) [25] are designed to detect the main regions of interest from a video frame and correct the spatial variations of the input data [26]. Similarly, in [27] a STN framework designed to capture the facial landmarks or facial visual saliency maps is proposed.

The facial expression recognition (FER) from video streams considers as single input a range of frames within a temporal analysis window. By using both textural and temporal information, it makes it possible to encode more stable expressions. In [28], authors use a C3D network, which consists in using 3D convolutional kernels with shared weights along the time axis instead of traditional 2D kernels. The network has been widely used for dynamic FER in [29-32]. Abbasnejad *et al.* [29] propose to deal with the lack of available data by generating large scale synthetic 3D faces. In [30], a late-fusion method that combines RNN with a C3D network is proposed. Ouyang *et al.* [31] use deep network transfer learning for feature extraction, a spatial-temporal model to capture the dynamics and reinforcement learning as optimization of the fusion strategies. In [32], a 3D-CNN is used to learn the static and dynamic features from facial image sequences and extract high-level dynamic features from optical flow sequences.

Other approaches propose designing FER models that are trained on single images, carefully selected from a large training dataset [33]. The supervised and self-supervised learning methods proposed in [34] are designed to improve the classification accuracy of fine-grained and in-the-wild FER. The Visual Transformer with Feature Fusion (VTFF) introduced in [35] is able to identify emotions in wild environments under extreme conditions. The system focuses on challenging cases of faces with important occlusions or deformations, with different poses or affected by motion blur. The transformer uses two branches: an attentional selective fusion mechanism that leverages between feature maps and a second part that models the relation between visual words and global self-attention maps. The TransFER facial expressions recognition method introduced in [36] consists of a visual transformer with both global and local attention mechanisms, specifically designed to extract rich relation-aware representations between visual descriptors. In [37], a Graph Convolutional Network (GCN) framework is proposed, that exploits the dependencies between categorical and dimensional emotion recognition tasks. Abbasi *et al.* [38] construct a graph-based representation for facial expressions recognition of children and predict the subjects' emotional state by using the automatically detected action units. Finally, a video-based FER is introduced in [39] that uses the emotion-wheel information as an inductive bias to improve the level of descriptiveness of the embedding features.

Globally, when analyzing the various state-of-the-art FER methods, the following conclusions can be highlighted: (1). Because the expression intensity in a video sequence varies over time it is very important to distinguish between peak and non-peak video frames; (2). The deep RNN/3D-CNN networks are designed to encode the temporal dependencies between consecutive frames. However, the performance of such systems is barely satisfactory. The RNN are unable to capture the powerful convolutional features, while the 3D-CNNs are applied over short video clips and ignore the long-range dynamics. In addition, training such architecture is highly difficult, notably when video data is insufficient.

Let us now review the multi-modal emotion recognition frameworks.

2.3. Multimodal emotion recognition systems

Several studies apply multimodal audio-video analysis to predict emotions [33-35]. Nguyen *et al.* [33] introduce a novel method that integrates 3D-CNN and deep belief networks to effectively model spatial and temporal information presented in video and audio for emotion recognition. The method uses a feature-level fusion approach based on a bilinear pooling theory to combine visual and audio feature vectors. In [34], two distinct sets of acoustic and visual features are applied as input to a CNN extended with an RNN. Finally, an average fusion method is applied at the decision level. Kahou *et al.* [35] combine multiple deep neural networks for different data modalities (*i.e.,* a CNN for facial expression analysis, a deep belief network to capture the audio information, a deep autoencoder for human action recognition and a shallow network for human mouth detection) into different aggregation strategies to predict emotions.

The VAANet approach introduced in [36] proposes a CNN architecture with spatial, channel and temporal attention mechanisms to identify emotions in users' generated videos. Both visual and audio information are here jointly exploited. Emotions are predicted using a late-fusion strategy that concatenates the descriptors from each modality. However, the framework neglects the interaction between the various features involved. The AVER [37] system proposes capturing the correlation between the A-V data using transformers with attention mechanisms over short video clips. Wang *et al.* [38] introduced an end-to-end knowledge injectable deep neural network able to associate implicit contextual knowledge when processing explicit A-V information. An audio-visual deep learning algorithm based on transformers is introduced in [39]. The fusion of the two modalities is performed using a cross-modal attention layer that consists of a dot-product attention of the key and value matrices computed from one modality with the query matrix given by the opposite modality. In [40], a system that fuses textual, audio and visual information is proposed. The sentence semantics is here extracted using a transformer architecture. The visual descriptors are obtained using a CNN model extended with an attention mechanism, while the temporal dynamics of the audio and visual signals are estimated with the help of a LSTM model. Recently, in [41], an optimal multimodal emotion recognition model is proposed that fuses the A-V features at the model level.

Hu *et al.* introduced in [42] a graph-based dynamic fusion method able to aggregate the information from visual, audio, and textual modalities by exploring both unimodal and cross-modal interactions in a graph structure. The module reduces the redundancy and enhances the complementarity between the modalities by capturing the dynamics of contextual information in different semantic spaces. The MEmoBERT approach presented in [43] learns multimodal joint representations through self-supervised learning from self-collected, large-scale, unlabelled video data. A multimodal music emotion recognition method that jointly predicts the valence and arousal values by combining the audio, lyrics, track name, and artist of a given track is introduced in [44]. Recently, multiple transformer-based frameworks [45-47] have been proposed to fuse and enrich multimodal features from raw videos for the task of multi-label video emotion recognition. Specifically, in [45] the method takes raw video frames, audio signals, and text subtitles as inputs and passes the information from multiple modalities through a unified transformer architecture for learning a joint multimodal representation. Chen *et al.* [46] propose a key-sparse transformer designed to focus only on relevant emotion-related features and to eliminate redundant information that limits the system performance. In [47], a three-branch transformer for video, audio and audio-video modalities is proposed. The experimental evaluation performed on two publicly available datasets demonstrates the powerfulness of multimodal concatenation of audio and video features that outperform single modalities.

In a general manner, the state-of-the-art analysis highlights the following conclusions: (1). Multimodal emotion recognition methods demonstrate superior performances compared to unimodal ones. Various fusion strategies

between different sources of information have been evaluated, including decision-level, feature-level, or model-level fusion. The model-level fusion returns the best results [41] because it leverages the complementary nature of the feature representations involved and extracts the cross-modal interaction between the audio-video channels. (2). Not all information from a video sequence is equally important for emotion recognition. In an image sequence, relevant emotional cues appear only in certain video frames. Similarly, in the audio stream not every word in the sequence contributes equally to the expressed emotion. (3). Most methods employ a two-stage shallow pipeline, which consists in extracting audio and visual features independently. They are thus agnostic to the complementary information between different A-V modalities. (4). To the best of our knowledge none of the state-of-the-art methods consider the correlation between various emotion classes, such as the emotions polarity defined by the Mikel's wheel [42].

In this paper, we propose a cross modal A-V fusion framework with double attention and deep metric learning that addresses the above problems for recognizing emotions, without requiring any auxiliary data except the initial pre-training of the various CNN architectures involved. We start by dividing each video stream into segments (*snippets*) containing a fixed number of keyframes that are further applied as input to a 3D-CNN architecture. The backbone ResNet 3D [43] network generates low-level feature representations that capture the spatial-temporal dynamics of the visual information. The audio module extracts low-level descriptors (*i.e.*, image spectrograms) from the corresponding audio stream. As for the visual representation, we start by dividing each image spectrogram into a set of audio samples that are fed as input to ResNet18 [44] CNN architecture. The audio and visual features are passed through different self-attention mechanisms designed to capture the intra-modal characteristic and determine the influence of the respective audio and visual descriptors to the final representation. In order to capture the inter-modal correlation between A-V information, a cross-attention method is proposed, which identifies salient relationships across the two modalities. Finally, the discrete emotion is predicted by passing the weighted descriptors through a set of fully connected layers. Under this framework, we design a novel classification loss that integrates an emotional metric constraint guiding the attention generation mechanisms. The relations between the considered emotional categories are here modeled according to Mikel's wheel [42].

The following section describes the proposed approach and details the various modules involved.

3. Methodology

Fig. 1 illustrates the proposed architecture that involves two processing chains, respectively dedicated to the audio and visual information. For the visual stream, three different attention mechanisms are employed, including spatial, channel and temporal, while the audio branch contains solely the temporal attention. The features from both modalities are exploited into a cross-attention fusion system. The training is performed by minimizing the proposed emotional metric loss in an end-to-end manner.

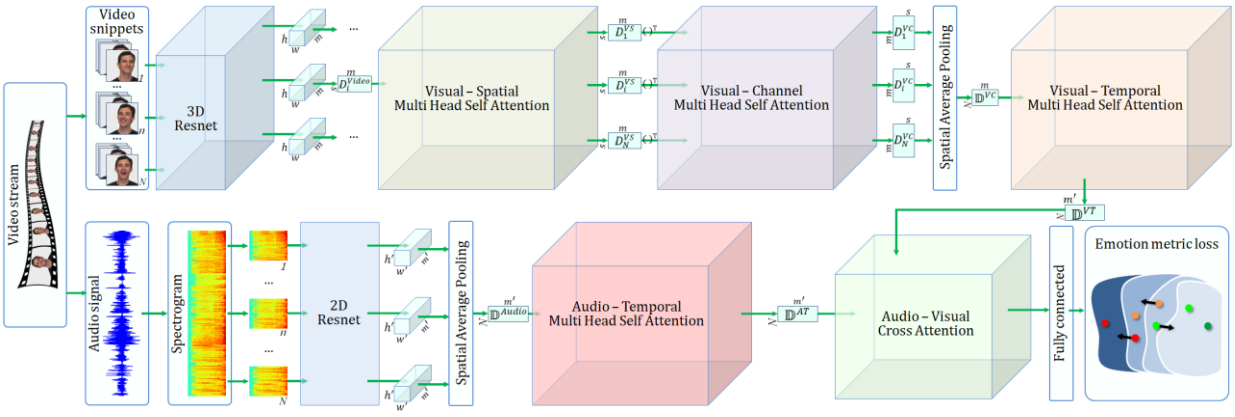

Fig. 1. Proposed architecture with main modules involved.

3.1. Visual stream analysis

The module extracts visual descriptors from variable length video documents (x^{Video}). A label y^{Video} is associated to each x^{Video}. The visual model is designed to work on short *snippets* uniformly sampled within the entire video stream, with no overlapping parts. Specifically, we divide each image sequence into segments denoted by $\{x_i^{Video}\}_{i=1}^{N}$, where *N* represents the total number of snippets. Each snippet contains a fixed number of *k* frames selected consecutively or not, depending on the length of the original input video.

We have decided to use a 3D-CNN model for the visual analysis because we argue that the image-based features extracted from 2D-CNN architectures are not directly suitable for emotion recognition in video streams due to the lack of temporal information. In contrast with existing still image datasets usually employed for machine learning purposes, the first video databases were of relatively small sizes (*i.e.*, a few thousands of training videos). However, with the development of Kinetics400 [62] or other large-scale datasets it becomes possible to reliably train 3D architectures and then perform finetuning for different tasks on smaller video sets. Nowadays, the 3D-CNNs have achieved significant advances and became the top performers on almost every video classification benchmark dataset [63].

We use the ResNet 3D [43] network as the backbone for the visual analysis module to generate high-level feature representations. The 3D ResNet101 performs 3D convolutions and 3D pooling. The CNN contains 101 layers grouped in 5 regions denoted conv1, conv2-x, conv3-x, conv4-x, conv5-x (that involve 3 successive convolution operations repeated x = 3, 4, 23 and 3 times, respectively). In our framework, we have removed the global average pooling layer at the end. The size of the convolutional kernels is 3 x 3 x 3 and the temporal stride is 1. Down-sampling of the inputs is performed by conv2-1, conv3-1, conv4-1 and conv5-1 layers, with a stride of 2. Each convolutional layer is followed by batch normalization and ReLU operations.

The 3D-CNN receives as input *N* snippets, independently processes them, and extracts visual descriptors from the last convolutional layer (conv5-3). For a sample x_i^{Video}, we denote by $D_i^{Video} \in \mathbb{R}^{h \times w \times m}$ the descriptor extracted at the output of the conv5 layer in the ResNet 3D architecture, where *m* is the total number of feature maps (*i.e.*, 1024 channels) and $h \times w$ is the spatial size (height and width) of the descriptor (*i.e.*, 16 x 16). By flattening $D_i^{Video} \in \mathbb{R}^{h \times w \times m}$ with respect to the height and width, we extract the output matrix $D_i^{Video} \in \mathbb{R}^{s \times m}$, with $s = h \times w$.

Estimating emotions by assigning equal importance to the visual feature maps extracted from 3D-CNN models may lead to sub-optimal prediction because: (1). Some image regions have reduced relevance (*e.g.*, the facial areas are more important when compared to the spatial context); (2). Only some feature maps carry discriminative semantic attributes necessary to identify emotions; (3). Not all video frames are equally important (*i.e.*, emotions can be predicted by using some keyframes depicting the peak of emotion). The framework proposed in this paper notably aims at overcoming such limitations and difficulties by applying three different attention mechanisms (spatial, channel-wise and temporal).

Spatial attention model: One important property of the human visual system is its capacity to obtain useful information with limited processing resources. Thus, humans are not processing the whole scene at once, but selectively focus on salient parts of images to capture the visual structure. Within the context of computer vision tasks, the process of spatial attention makes it possible to focus on regions that carry discriminative information within each feature map and thus to enhance the relevance/influence of such regions within the visual descriptor. Attention can be considered as a dynamic feature extraction mechanism that combines contextual fixation over time.

We have developed a visual spatial (VS) multi-head self-attention module [45] to automatically explore the relevance of the video frame regions in order to recognize emotions (Fig. 2). Each head of attention requires three matrices: $Query_h^{VS} \in \mathbb{R}^{s \times m/H}$, $Key_h^{VS} \in \mathbb{R}^{s \times m/H}$ and $Value_h^{VS} \in \mathbb{R}^{s \times m/H}$. The projection on each head of attention $h \in \{1, \dots, H\}$ is determined as described in Eq. 1:

$$Query_h^{VS} = D_i^{Video}\, W_h^Q \;;\; Key_h^{VS} = D_i^{Video} W_h^K \;;\; Value_h^{VS} = D_i^{Video}\, W_h^V, \tag{1}$$

where *H* represents the number of attention heads, while $W_h^Q \in \mathbb{R}^{m \times m/H}$, $W_h^K \in \mathbb{R}^{m \times m/H}$ and $W_h^V \in \mathbb{R}^{m \times m/H}$ are three learnable parameter matrices. The spatial attention mechanism for a given attention head can be computed as:

$$Attention_h^{VS}\left(Query_h^{VS}, Key_h^{VS}, Value_h^{VS}\right) = softmax\left(\frac{Query_h^{VS}\left(Key_h^{VS}\right)^{\mathrm{T}}}{\sqrt{m/H}}\right) Value_h^{VS}, \tag{2}$$

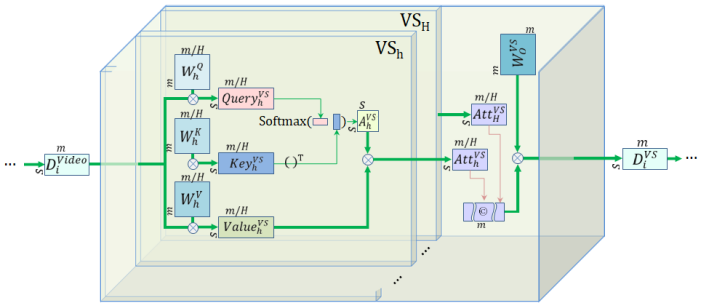

Fig. 2. The VS self-attention module.

For one attention head the $Query_h^{VS}$ matrix can be interpreted as the set of embeddings on which we want to apply the attention mechanism, while the Key_h^{VS} represents the set of features against which we compute the attention. As the result of the dot multiplication and softmax(·) the system determines a set of weights that are further multiplied with the $Value_h^{VS}$ matrix to determine the spatial attention for one attention head.

Instead of performing a single attention function it is beneficial to linearly project the queries, keys, and values H times with different, learnable linear projections. On the projected version we compute the spatial attention in parallel. For each corresponding video snippet x_i^{Video}, the results of all H attention heads are concatenated and once again projected using an extra parameter matrix $W_O^{VS} \in \mathbb{R}^{m \times m}$:

$$D_i^{VS} = Concat(\, Attention_1^{VS}, \dots, Attention_H^{VS}) W_O^{VS}, \tag{3}$$

where $D_i^{VS} \in \mathbb{R}^{s \times m}$ represents the weighted spatial visual descriptor returned at output.

The channel-wise attention model: We introduce the channel attention by exploiting the relevance of various visual feature maps extracted from the convolutional layers. As each channel can be considered as a feature detector, the channel attention is designed to focus on "what" is meaningful within a given input image. The core idea is to automatically learn a set of weights for each channel of the feature map, in order to assign larger weights to the visual descriptors that contain more important information, while ensuring smaller weights for invalid or less discriminant feature maps.

Assuming that each channel (*i.e.*, a column in the spatial visual descriptor D_i^{VS} matrix) corresponds to the responsive activation of a convolutional kernel in the last layer of the 3D-CNN, the channel-wise attention can be interpreted as the process of selecting the relevant semantic attributes from the visual descriptors. Within this context, we introduce a channel-wise (VC) multi-head self-attention module that automatically assigns higher scores to relevant feature maps. To generate the channel-wise attention, we first transpose D_i^{VS} to $(D_i^{VS})^{\mathrm{T}} \in \mathbb{R}^{m \times s}$. For each head of attention, we used the $Query_h^{VC} \in \mathbb{R}^{m \times s/H}$, $Key_h^{VC} \in \mathbb{R}^{m \times s/H}$ and $Value_h^{VC} \in \mathbb{R}^{m \times s/H}$ matrices and computed the projection on each sub-space $h \in \{1, \dots, H\}$ as described in Eq. 4:

$$Query_h^{VC} = (D_i^{VS})^{\mathrm{T}} W_h^Q \;;\; Key_h^{VC} = (D_i^{VS})^{\mathrm{T}} W_h^K \;;\; Value_h^{VC} = (D_i^{VS})^{\mathrm{T}} W_h^V, \tag{4}$$

where $W_h^Q \in \mathbb{R}^{s \times s/H}$, $W_h^K \in \mathbb{R}^{s \times s/H}$ and $W_h^V \mathbb{R}^{s \times s/H}$ are three learnable parameter matrices.

The channel attention can be mathematically expressed as presented in Eq. 5.

$$Attention_h^{VC}(Query_h^{VC}, Key_h^{VC}, Value_h^{VC}) = softmax\left(\frac{Query_h^{VC}(Key_h^{VC})^{\mathrm{T}}}{\sqrt{s/H}}\right) Value_h^{VC}, \tag{5}$$

Finally, in order to determine the output of the module ($D_i^{VC} \in \mathbb{R}^{m \times s}$), the results from all attention heads (H) are concatenated and projected using the matrix $W_O^{VC} \in \mathbb{R}^{s \times s}$.

The temporal attention model: For a video, the discriminability of each frame to recognize emotions is obviously different. Only a limited number of frames contain discriminative information and can be directly used to convey emotions, while the rest correspond to background/contextual elements or preparatory stages. Treating each frame with equal importance may mislead the classifier and thus be responsible of wrong predictions. Based on such

observation we design a temporal attention mechanism able to automatically focus only on the video snippets that contain relevant keyframes.

The first stage of the visual temporal (VT) multi-head self-attention module is to apply the spatial average pooling over the visual descriptor $\{D_i^{VC}\}_{i=1}^{N}$ returned by the channel attention module and reshape it to: $\mathbb{D}^{VC} = [d_{v1}, \dots, d_{vN}] \in \mathbb{R}^{N \times m}$. We denote by d_{vi} the visual descriptor associated to the i^{th} video snippet. In this context, the temporal attention can be defined as presented below:

$$Attention_h^{VT}(Query_h^{VT}, Key_h^{VT}, Value_h^{VT}) = softmax\left(\frac{Query_h^{VT}(Key_h^{VT})^{\mathrm{T}}}{\sqrt{m/H}}\right) Value_h^{VT}, \quad (6)$$

The Query, Key and Value $\in \mathbb{R}^{N \times m/H}$ matrices are computed using the $\mathbb{D}^{VC}$ visual features and the associated learnable parameter matrices (*cf.* Eq. 1). In order to determine the visual descriptor returned by the module ($\mathbb{D}^{VT} \in \mathbb{R}^{N \times m'}$), the outputs from all the attention heads (*H*) are concatenated and projected using the matrix $W_O^{VT} \in \mathbb{R}^{m \times m'}$.

3.2. Audio stream analysis

Emotions play an important part in our daily life.? They can reflect psychological and physical states, but also have a vital role in human communication and cognition. The physiological signals such as electroencephalogram (EEG), temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiration (RSP)) reflect responses to the central nervous system (CNS) and to the autonomic somatic nervous systems (ANS) of the human body, in which the human states change according to Connon's theory [64]. Various researchers have tried to establish a stand and fixed relationship between the psychological signals, features and the classifiers. However, it has been demonstrated that it is difficult to precisely reflect emotional changes using physiological signals in both research and real applications [65]. Based on such observations, in our work we have decided to use the audio signal that is still one of the primary contact-free modalities used to convey emotions.

The audio features are complementary to the visual descriptors and contain elements of information that can be used to characterize the affective, psychological state of an individual. To describe the audio signal, we have considered the most well-known audio representation that can be applied as input to a CNN system: the image spectrogram [46].

The module extracts audio descriptors from variable length audio documents (x^{Audio}). A label y^{Audio} is associated to each x^{Audio}. As for the visual representation, we start by dividing each image spectrogram into *N* audio samples $\{x_i^{Audio}\}_{i=1}^{N}$ that are fed as input to ResNet18 [44] CNN architecture. The 2D ResNet contains 18 layers grouped in 5 regions denoted conv1, conv2-x, conv3-x, conv4-x, conv5-x (that involve 2 successive convolution operations repeated x = 2 times). In our framework we have removed the global average pooling layer at the end. The size of the convolutional kernels is 3 x 3. The down-sampling of the inputs is performed by conv2-1, conv3-1, conv4-1 and conv5-1 layers, with a stride of 2. Each convolutional layer is followed by batch normalization and ReLU.

We have used ResNet18, a 2D-CNN architecture designed for images where the convolutional kernels share the same weights across the horizontal and vertical axis. This is based on the translational invariance of the visual features, which means that the visual descriptors of an object are the same, no matter the location of the object within the image. For spectrogram images, this property remains true if the objects are shifted in the *x* dimension (time), but not shifted in the *y* dimension (frequency). The 2D-CNN framework has been modified in order to be adapted for spectrograms inputs. It independently processes all spectrograms segments and extracts audio descriptors from the last convolutional layer (conv5). For a sample x_i^{Audio}, we denote by $D_i^{Audio} \in \mathbb{R}^{h' \times w' \times m'}$ the descriptor extracted at the output of the conv5 layer in the ResNet architecture, where m' is the total number of feature maps and $h' \times w'$ is the spatial size of the descriptor (height and width). Then, the audio temporal (AT) multi-head self-attention module explores the influence of different audio segments. We apply the spatial average pooling over $\{D_i^{Audio}\}_{i=1}^{N}$. and reshape it to a global feature representation $\mathbb{D}^{Audio} = [d_{a1}, \dots, d_{aN}] \in \mathbb{R}^{N \times m'}$, where d_{ai} represents the audio descriptor associated to the i^{th} image spectrogram segment.

Finally, the temporal attention is computed similarly as for the visual module (Eq. 6), which aims at quantifying the influence of the information included in different audio segments for the emotion recognition process. The output of the module is given by the $\mathbb{D}^{AT} \in \mathbb{R}^{N \times m'}$.

3.3. Cross-modal attention fusion

The audio-video fusion can be performed into three major stages: early, late or fusion at the level of the model. In early fusion [47], [48] the features from different modalities are concatenated after extraction in order to obtain a joint representation that is fed into a single classifier to predict the final outputs. Although such an approach allows the direct interaction between the modalities, it fails to leverage the inter-modal relationships and may suffer from data sparseness. So, the improvement in performance given by the concatenation of both modalities is only marginal. In contrast to feature-level fusion, in decision (late) fusion [49], [50] the classifiers are trained end-to-end independently, and the prediction outputs are combined to obtain the final classification results. Although decision-level fusion is easy to implement and does not require any additional training, such systems neglect the mutual relations and the correlation between the A-V modalities. The model-level fusion [45], [51], [52] uses the deep learning architectures to leverage the complementary nature of the various features involved and to extract the cross-modal interaction between the A-V channels.

The A-V fusion mechanism introduced in this paper is a model-based method designed to encode the inter-modal correlation, while preserving the relevant and distinctive intra-modal information (Fig. 3).

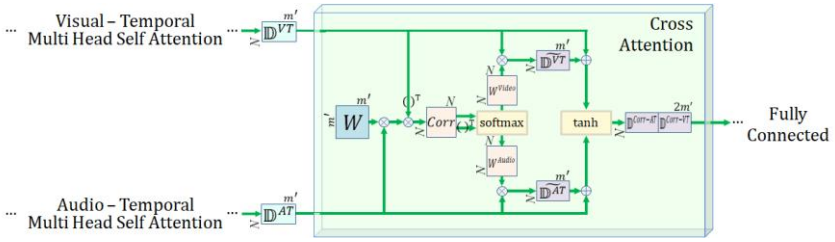

Fig. 3. The cross-modal attention fusion module.

The cross-modal attention fusion module receives as input the visual and the audio features returned at the output of the temporal attention modules presented in Section 3.1 and Section 3.2. The visual features are denoted with $\mathbb{D}^{VT} \in \mathbb{R}^{N \times m'}$, while the audio features with $\mathbb{D}^{AT} \in \mathbb{R}^{N \times m'}$, where N represents the total number of snippets selected from video/audio stream and m' is the size of the output feature descriptor (m'=512). It can be observed that both modalities use the same number of $N = 6$ samples and return feature descriptors of equal sizes (m').

To reliably fuse the two visual and audio modalities, we have developed a cross-attention mechanism, where features are separately learned for each modality under the constraints of the other modality. To reduce the gap of heterogeneity between the audio and visual modalities, the inter-modal relevance is determined using the W matrix $\left(W \in \mathbb{R}^{m' \times m'}\right)$, whose weights are learned during training as follows:

$$Corr = \mathbb{D}^{AT}\, W\, (\mathbb{D}^{VT})^{\mathrm{T}}, \tag{7}$$

where W represents the cross-correlation weights among the A-V features and m' is the dimension of the feature vectors from both modalities. In the resulting inter-modal correlation matrix $Corr \in \mathbb{R}^{N \times N}$, high coefficient values indicate that the corresponding pairs of A-V segments are highly correlated. For each modality the attention weights (W^{Audio} and W^{Video}) are computed as the column-wise softmax of $Corr$ and $Corr^{T}$, respectively. Then, for each modality, the attention weights are used to assign different relevance scores to the features. Formally, the cross-attention weights for the visual ($\widetilde{\mathbb{D}^{VT}}$) and audio ($\widetilde{\mathbb{D}^{AT}}$) modalities are computed as:

$$\widetilde{\mathbb{D}^{AT}} = W^{Audio}\mathbb{D}^{AT} \text{ and } \widetilde{\mathbb{D}^{VT}} = W^{Video}\mathbb{D}^{VT}, \tag{8}$$

We have used the skip connection to adjust the attention maps by taking into consideration the information retrieved from the different modality:

$$\mathbb{D}^{Corr-AT} = tanh\left(\mathbb{D}^{AT} + \widetilde{\mathbb{D}^{AT}}\right), \tag{9}$$

$$\mathbb{D}^{Corr-VT} = tanh\left(\mathbb{D}^{VT} + \widetilde{\mathbb{D}^{VT}}\right), \tag{10}$$

where $\tanh(\cdot)$ denotes the hyperbolic tangent activation function.

The final data representation is obtained by the direct concatenation of the audio and visual descriptors:

$$\mathbb{D}^{Corr-AV} = [\mathbb{D}^{Corr-VT}; \mathbb{D}^{Corr-AT}], \tag{11}$$

The $\mathbb{D}^{Corr-AV}$ features are fed as input to the fully connected layers. Applying the cross-correlation stage allows to bring the audio and visual embedding closer, while the dense skip connection enforces the modality-specific information. By using the cross-modal attention module during the multiple stages of the training process we are able to progressively learn optimal embeddings. Under such constraints, the framework can learn the right amount of compatibility between the two embeddings considered and thus preserve the intra-class information, while optimizing the objective function.

The framework adopts an emotion metric that enables to generate polarity-preserved attention maps, described in the following section.

3.4. Emotion metric loss

In order to define the distance between emotions we have considered the 2D valence-arousal model described by the Mikel's [42], wheel of emotions (Fig. 4a). By taking into consideration the discrete set of emotions identified by Eckman and Friesen in [2] and [3] we can establish three classes of emotions with positive polarity (*i.e.*, happy, calm and surprise) and four emotions categories with negative polarity (*i.e.*, sad, fear, anger and disgust). The neutral state has zero values for the valence and arousal, being located in the central of the wheel. The distance between different video clips depicting emotions is arguable subjective, but the general relationship is clear and should be satisfied: video stream presenting emotions of the same polarity should be closer to each other while those of different polarity should be further apart.

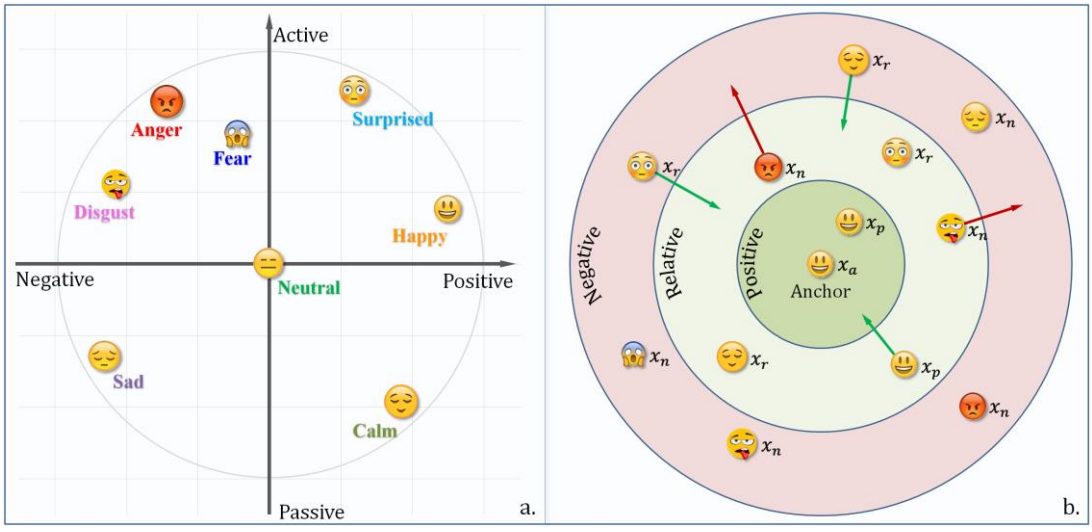

Fig. 4. Hierarchical relation between emotions: a. Emotion representation with respect to Mikel's wheel in the valence-arousal space; b. The proposed triplet loss function extended with an emotional constraint.

Let us first review the conventional triplet loss function in its mathematical form. Suppose that we are given a set of training examples, where $\boldsymbol{x} \in \mathbb{R}^D$ denotes the feature embedding and $\boldsymbol{y} \in \{1, .., K\}$ the corresponding label, with K

the total number of classes. At each training iteration, we sample a mini-batch of triplets: $\mathcal{T} = (x_a, x_p, x_n) \in \Gamma$, where Γ is the set training examples, that consists of an anchor point x_a, associated with a pair of positive instance x_p and a negative sample x_n, whose labels satisfy: $y_a = y_p \neq y_n$. The goal is to learn an embedding function to push away the negative example x_n from the anchor sample x_a by a distance margin α, compared with the distance between the anchor and the positive sample x_p:

$$\left\|x_a - x_p\right\|^2 + \alpha = \|x_a - x_n\|^2, \tag{12}$$

where $\|\cdot\|$ represents the Euclidian distance.

During training, in order to force the system to respect the triplet loss constraint described in Eq. 12, a hinge loss function is commonly used:

$$Loss = \sum_{i=1}^{N} \left[\left\|x_a - x_p\right\|^2 - \|x_a - x_n\|^2 + \alpha\right]_+, \tag{13}$$

where $[\cdot]_+ = max(0, \cdot)$ denotes the hinge function and N is the total number of samples from the training set.

However, directly optimizing the triplet loss function can lead to some videos to be incorrectly classified into categories that have opposite polarities. In addition, the natural relations between emotions are not considered. In our work, we propose to enrich the triplet loss function with an emotion constraint. The proposed principle is the following: as in the case of the traditional triplet loss approach, we ensure that an anchor example x_a is closer to all samples x_p belonging to the same emotion class. In addition, we introduce the notion of *related emotions*, which are, by definition, negative emotion categories that have the same polarity as the considered emotion. We then constrain the related categories to be closer to the considered anchor class than the hard negative examples defined as negative emotions with opposite polarity (Fig. 4b). The emotion loss is mathematically formulated as presented in the following equations:

$$\left\|x_a - x_p\right\|^2 + \alpha = \left\|x_a - x_{r_1}\right\|^2, \tag{14}$$

$$\left\|x_a - x_{r_2}\right\|^2 + \beta = \|x_a - x_n\|^2, \tag{15}$$

where $\alpha > 0$ and $\beta > 0$ denote the control parameters (margins) between the considered classes.

The emotion metric is learned by minimizing the following loss function:

$$Loss_{emotion} = \sum_{i=1}^{N} \left[\left\|x_a - x_p\right\|^2 - \left\|x_a - x_{r_1}\right\|^2 + \alpha\right]_+ + \sum_{i=1}^{N} \left[\left\|x_a - x_{r_2}\right\|^2 - \|x_a - x_n\|^2 + \beta\right]_+, \tag{16}$$

Let us note that in the case of the neutral category, no related emotion can be identified. For video samples depicting the neutral emotion the traditional triplet loss is used. Since the emotion metric loss can be computed within mini-batches of samples, we can train the entire framework effectively, in an end-to-end manner using off-the shelf-optimizers to minimize the loss function in Eq. 16.

4. Experiments and results

To evaluate the effectiveness of the proposed method within the context of discrete emotion recognition, we have conducted extensive experiments on two publicly available datasets: RAVDESS [4] and CREMA-D [5]. Both datasets contain discrete emotions, performed at two levels of intensity (normal and strong), and recorded using both audio and visual channels. In addition, the ground truth provided by the human annotation is also available.

4.1. Evaluation datasets

The RAVDESS dataset [4] is a multimodal, gender-balanced database consisting of 24 professional actors, vocalizing two lexically matched statements in English, with North American accent. The dataset contains 1440 videos

with eight basic emotions: surprise, anger, fear, neutral, calm, happy, sad, and disgust. There are two different levels of intensity (normal and strong) considered for each expression of emotion, except for the neutral class that is represented only at the normal level. All types of recordings are available: audio only, video only and both audio and video. The recordings have an average duration of 3.82 ± 0.34 seconds. In our experiments, we have considered exclusively video streams with both audio and visual information because our goal is to perform multimodal emotion recognition.

The CREMA-D dataset [5] consists of facial and vocal emotional expressions in sentences with 6 basic emotion labels: happy, sad, anger, fear, disgust and neutral. The dataset consists of 7440 video clips spoken by 91 actors (43 females and 48 males) accounting a total of 3.5 hours. The age of the actors lays in range of 20 to 74 years with diverse ethnic background including: African American, Caucasian, Asian, Hispanic, and not specified. Actors read from a pool of 12 sentences to generate the discrete emotion dataset at four intensity levels: low, medium, high, and unspecified. The recorded video clips have been labelled by multiple human annotators in three modalities: audio only, video only and audio-visual. The video clips have an average length of 3.63 ± 0.53 seconds. In our experiments, we have considered exclusively the audio-video modality. The human recognition rate of the intended emotions for the audio-video stream is 63.6%.

Despite the increasing number of publications that use RAVDESS [4] or CREMA-D [5] datasets, there is a lack of a common evaluation framework. To the very best of our knowledge, there is no technical evaluation baseline for the considered datasets, which makes it very difficult to compare the state-of-the-art techniques. For example, Atila *et al.* [19] reported an accuracy score of 96.1% on the RAVDESS database, but the actors' distribution in folds is not specified. In addition, it is not clear if the same actor takes part in the training, but also in the testing sets. A different setup for evaluation is proposed by Pepino *et al.* [10], where 20 actors are used for training, two for validation and other two for testing. However, nothing is said about how the two actors in the testing dataset have been selected. The reported performance is dependent on the two single actors' evaluation, which may or may not reflect the performance scores at the level of the entire dataset. In addition, the authors classify only seven (out of eight) emotions from the RAVDESS database.

For all these reasons, in order to perform a fair evaluation, we have divided the data into training, validation and testing sets with all the samples from each speaker belonging to a particular set only and we have used all the emotion categories existent in the evaluation dataset. We have adopted a 10-fold cross validation with a split driven by the actors that are expressing the utterances. Thus, no overlap is allowed between the subjects' clips: a video document is either in the test, validation or in the train datasets. The targeted percentages between train, validation and test are of 80%, 10% and 10%, respectively. The results that are reported in the following sections represent the average values for the 10-folds cross validation.

4.2. Implementation details

For the *visual modality*, the system is based on the state-of-the-art 3D ResNet101 [43] architecture that receives as input aligned face instances with the resolution of 224 x 224 pixels. The 3D-CNN is initialized with a set of weights pre-trained on ImageNet [52] and Kinetics [53]. The video streams are divided into a fixed number of shorter snippets (N = 6). Each snippet contains 16 frames uniformly sampled. In total, 96 keyframes are selected from each video stream for further analysis. The number of frames between the selected keyframes varies with respect to the length of the input video (between 3.1 and 4.16 seconds), thus ranging within $[0, 2]$ interval. In order to increase the variety of data, we have performed data augmentation on the training dataset by randomly cropping, flipping and brightness adjusting the input frames. For attention mechanisms we have considered H = 8 attention heads.

For the *audio module* we fed as input to the 2D ResNet-18 [44], image spectrograms extracted from the speech signals. We have applied the Fast Fourier Transform with 256 frequency components and we have generated image spectrograms using a Hamming sliding window of 32 ms with a step of 10 ms. As for the visual representation, the image spectrogram is divided into N =6 shorter parts of one second of audio content. In order to increase the system robustness and reduce the training time, the mean and variance normalization are performed over the image sequences. Specifically, each batch of elements is normalized to have zero mean and unit variance. This scheme introduces additional randomness in the network, since the output of a unit depends on the mini-batch statistics, as well as the input sample. In addition, the batch normalization helps reducing the training convergence time and the validation loss

to half. The 2D-ResNet [44] receives as input images with the resolution of 224 x 224 pixels and is initialized with the sets of weights pre-trained on ImageNet [52].

For the *audio-video fusion* module, we have used the hyperbolic tangent as activation function. The weights of the matrices in the cross-attention module are initialized with the Xavier method [55]. We have used the Xavier initialization because with a standard weight initialization we have observed a sequentially occurring saturation phenomenon that is propagating up in the network, while the He initialization method [61] works better for layers with ReLU activation functions.

The initial learning rate of the network was set to 1e-4 and a momentum of 0.9 is used with the Adam optimizer [54]. Because of the hardware limitations and memory constraints we have chosen a batch size of 16. To avoid overfitting, we have implemented an early stopping strategy to finish the training process: when the accuracy score does not improve in 20 epochs the training stops.

4.3. Ablation study

We have conducted ablation studies to verify the individual impact of the various components involved on the overall performance. More precisely, the main components under evaluation are the following: the visual attention mechanisms, the temporal attention mechanism dedicated to the audio modality, the cross-attention fusion strategy and finally the emotional-constrained loss function.

First, we have used the triplet loss, and have investigated the influence of the various attention modules involved: the visual spatial (VS), the visual channel-wise (VC), the visual temporal (VT), the audio temporal (AT) and the audio-video fusion. The emotion recognition accuracy rates of each category for the two datasets considered are summarized in Table 1 and 2, respectively. For all the testing scenarios, the framework architecture ends with one fully connected layer having the number of neurons equal to the total number of emotion classes. In all these cases, the traditional triplet loss function is used to train the network.

After analyzing the experimental results gathered in Table 1 and Table 2, the following conclusions can be derived:

1. The lowest accuracy scores are returned by the audio modality, with 76% and 62%, respectively, for the RAVDESS and CREMA-D datasets. The results can be explained by the nature of the audio data representation that depends on multiple factors such as sample duration, speaker accent or gender. However, if we analyze the emotion recognition rate provided by human observers on CREMA-D dataset using only the audio modality (49%) with the results obtained by the proposed framework it can be observed that our method outperforms human perception with more than 10%.

2. The visual modality, even when using only the spatial attention, significantly outperforms the audio modality, which is understandable because all the videos from both datasets are recorded in a specific environment setting with little variation of the subject position.

3. Adding attention (channel-wise and temporal) leads to performance gains, which demonstrate that the channel-wise and temporal attention contribute to the video emotion recognition framework.

4. The best results (with accuracy scores of 87.85% and 81.71% on the RAVDESS and CREMA-D datasets, respectively) are obtained by combining the audio and visual features (*cf.* Section 3.3).

Table 1. Ablation study of the different attention mechanism involved in the proposed framework on the RAVDESS dataset.

Method	Surprise %	Angry %	Fearful %	Neutral %	Calm %	Happy %	Sad %	Disgust %	Average %
AT	88.75	88.13	68.75	81.25	91.88	62.52	66.88	65.63	76.42
VS	79.38	86.88	75.63	65.02	83.13	86.56	59.38	78.13	77.54
VS + VC	82.51	88.13	77.52	67.51	84.38	88.75	60.01	78.75	79.17
VS + VC + VT	82.02	87.51	80.01	75.25	84.25	92.51	66.88	84.38	82.02
AT + VS + VC + VT	86.88	93.13	86.88	86.75	91.01	92.75	78.25	86.63	87.85

Table 2. Ablation study of the different attention mechanism involved in the proposed framework on the CREMA-D dataset.

Method	Anger	Fear	Neutral	Happy	Sad	Disgust	Average

	%	%	%	%	%	%	%
AT	73.75	50.63	72.63	49.63	65.63	64.38	62.77
VS	84.94	64.38	76.25	90.63	63.13	77.51	76.14
VS + VC	85.69	62.56	77.51	91.88	65.63	79.38	77.12
VS + VC + VT	87.94	67.44	80.44	93.44	65.25	82.56	79.51
AT + VS + VC + VT	88.81	70.25	84.63	93.69	67.94	84.94	81.71

In addition, it can be observed that emotion recognition is a highly challenging process because emotions are complex or compound and are unlikely to be fully learned through only one modality. In contrast, emotions like happiness have clear markers in at least one modality (*e.g.*, facial smile) and can therefore be learned with reasonable accuracy through visual classifiers. The facial expressions conveyed happiness the clearest, while vocal features conveyed anger better than other emotions. From our experiments, we have observed that in the context of emotion recognition, the gender of the speaker does not influence the performance scores obtained by the multimodal system.

Secondly, we evaluated the effectiveness of the proposed emotion constraint loss by comparing it with the results obtained for the traditional triplet loss function. Table 3 presents the average accuracy rates obtained when all the attention mechanisms are involved on both audio and visual modalities.

Table 3. Performance comparison of triplet loss function and the proposed emotion constraint loss

Attentions	Loss	RAVDESS	CREMA-D
AT + VS + VC + VT	Triplet loss	87.85%	81.71%
AT + VS + VC + VT	Emotion constraint loss	89.25%	84.57%

The results show that the introduction of the emotional constraint offers an average improvement of about 2%. This demonstrates the interest of introducing emotion relationships as prior knowledge to the classification task. Fig. 5 presents the confusion matrices obtained on the two considered datasets. As it can be observed for both datasets the higher recognition scores are obtained for the angry and happiness emotion categories. The experimental results demonstrate the powerfulness of analyzing both audio and visual features that can offer complementary information when performing emotion classification.

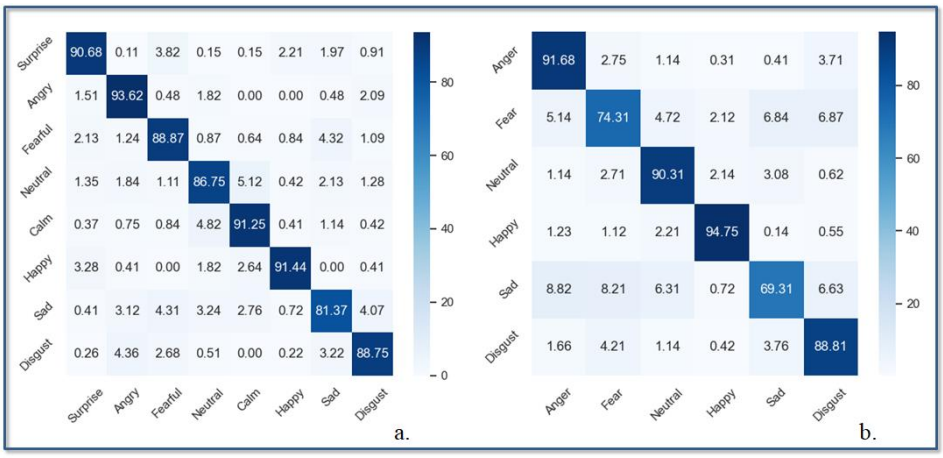

Fig. 5. The confusion matrixes on the evaluation datasets: (a). RAVDESS and (b). CREMA-D

4.4. System complexity evaluation

We have considered the following complexity metrics to quantify the resources required by the overall framework architecture:

1. *The training time* (min/epoch) – represents the amount of time, expressed in minutes, required to train the entire system for a single epoch. The reported value includes any time required within the low-level processing stages, the forward and the backward propagation passes of all batches.

2. *Inference time* (ms/sample) – represents the amount of time, expressed in milliseconds (ms), required to perform the inference of a single snippet containing 16 frames. However, we need to highlight that a fixed number of N =6 snippets are processed in parallel, by the 3D-CNN/2D-CNN architectures in order to extract the output features.

3. *The maximum GPU utilization* (GB) – we have evaluated the maximum GPU utilization, expressed in GBytes (GB) during training in order to determine the optimal hardware requirements.

For a fair comparison, all experiments have been conducted on a working station with Intel i7-8700K CPU at 3.70GHz, 32GB of RAM and Nvidia 3090Ti GPU. In order to minimize the impact of various randomly selected seeds involved within the system, we have conducted the experiments 10 times and computed the mean and standard deviation of the evaluation metrics. To avoid overcrowding the results, in Table 4 we report the mean values of the considered metrics.

Table 4. The proposed system complexity evaluation

Method	Training time (min/epochs)	Inference time (ms/sample)	Peak GPU usage (GB)
AT + VS + VC + VT with emotional constrain loss	95	235	12

Let us observe that the prediction time for a sample of 6 snippets with 16 frames is inferior to 250ms. Within this context, we can conclude that the proposed method can operate in real-time and is able to return a prediction for approximatively 4 seconds of video stream in less than 250ms.

4.5. Comparison with the state-of-the-art

We have compared the proposed framework with the following recent (published in the last here years) state of the art emotion recognition methods: Ghaleb *et al.* [37], Su *et al.* [56], Fu *et al.* [57], Jimenez *et al.* [27], Chang *et al.* [59], Jimenez *et al.* [27] and Middya *et al.* [41]. All the methods retained for comparative evaluation report results on the the two datasets considered (RAVDESS and CREMA-D). The experimental results obtained are summarized in Table 5.

Table 5. Comparison between the proposed framework and several state-of-the-art results on the RAVDESS and CREMA-D datasets

Method	Publication year	Accuracy on RAVDESS	Accuracy on CREMA-D
Ghaleb *et al.* [37]	2020	79.00%	74.00%
Su *et al.* [56]	2020	74.86%	-
Fu *et al.* [57]	2021	75.76%	-
Jimenez *et al.* [27]	2021	80.08%	-
Middya *et al.* [41]	2022	86.00%	-
Goncalves *et al.* [58	2022	-	71.7%
Chang *et al.* [59]	2021	-	83.15%
Proposed method	-	**89.25%**	**84.57%**

From the experimental results reported in Table 5 we can observe that the proposed framework outperforms the other models, with average accuracy scores of 89.25% on RAVDESS and 84.57% on CREMA-D dataset, respectively.

This behavior can be explained by the complexity of our method that exploits spatial, channel and temporal attention models to determine the intra-modal characteristics and the cross-attention fusion strategy to determine the correlation between the audio and video feature representation. In addition, the emotion constraint loss increases the system robustness to various types of noises existent in the video stream. In order to validate such an assumption, we have conducted an interpretability study, based on the visualization of the various attention maps involved and presented in the following section.

4.6. Interpretability: visualization of features

To investigate the interpretability of our framework, Fig. 6 presents the data distribution on the CREMA-D dataset. We have used the t-SNE embedding and generated a 2D representation of the sample feature distribution. For visualization purposes, in Fig. 6a we represent the samples in the 2D space when using the proposed framework with all the attention units involved from both modalities and adopting for the last fully connected layer the softmax loss function. We can observe that the classes are overlapping for most of the categories. Fig. 6b illustrates the feature space after applying the triplet loss function. In this case, it can be observed that the triplet loss function improves the feature separation into the novel sub-space. However, some clusters overlapping are still present, especially for emotion classes belonging to different polarities. Fig. 6c presents the same data distribution in the feature space, but when incorporating the emotion constraint loss introduced in Section 3.4. As it can be observed, our framework can separate the emotion classes more effectively. For the purpose of generalization, we have conducted the same set of experiments on RAVDESS database. From Fig. 7 we can observe a similar behavior for the data distribution.

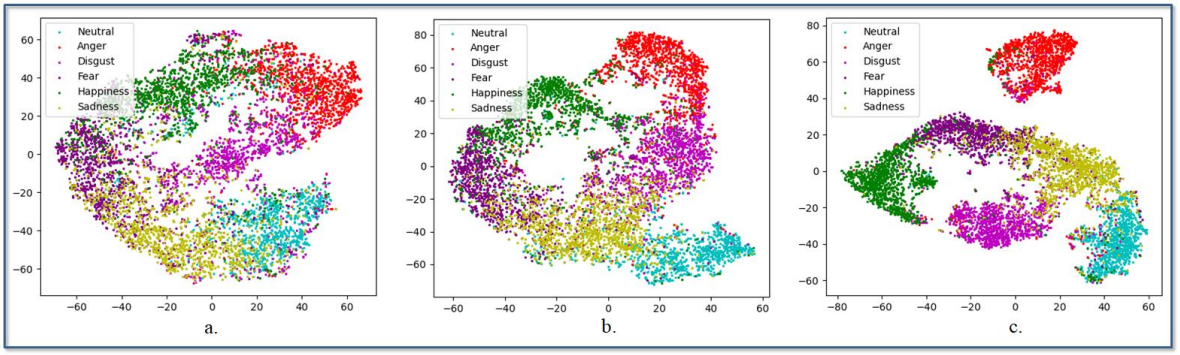

Fig. 6. Visualization of the audio-video feature embedding using t-SNE on the CREMA-D dataset: a. The feature space representation when using the softmax loss function; b. The feature space representation for triplet loss function; c. The feature space representation for triplet loss function extended with an emotional constraint.

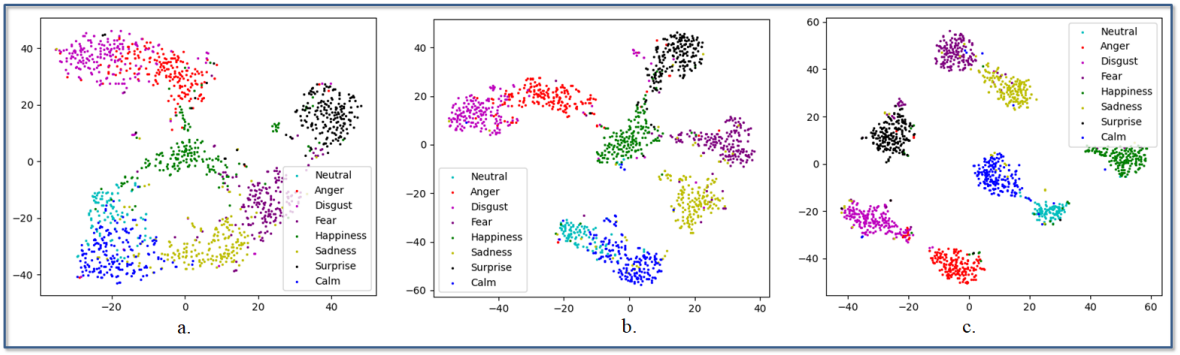

Fig. 7. Visualization of the audio-video feature embedding using t-SNE on the RAVDESS dataset: a. The feature space representation when using the softmax loss function; b. The feature space representation for triplet loss function; c. The feature space representation for triplet loss function extended with an emotional constraint.

Furthermore, we have studied the interpretability of the various attention mechanisms involved in the framework. To this purpose, we have exploited the Grab-Cam algorithm [60] that allows the visualization of the spatial attention mechanism on each attention head considered (Fig. 8 and Fig. 9). The final spatial attention, together with the visual temporal attention generated by our model is illustrated on various video snippets in Fig. 10 and Fig. 11. It can be observed that the proposed framework can successfully assign higher relevance scores not only to the discriminative frames, but also to various salient regions in the corresponding frames.

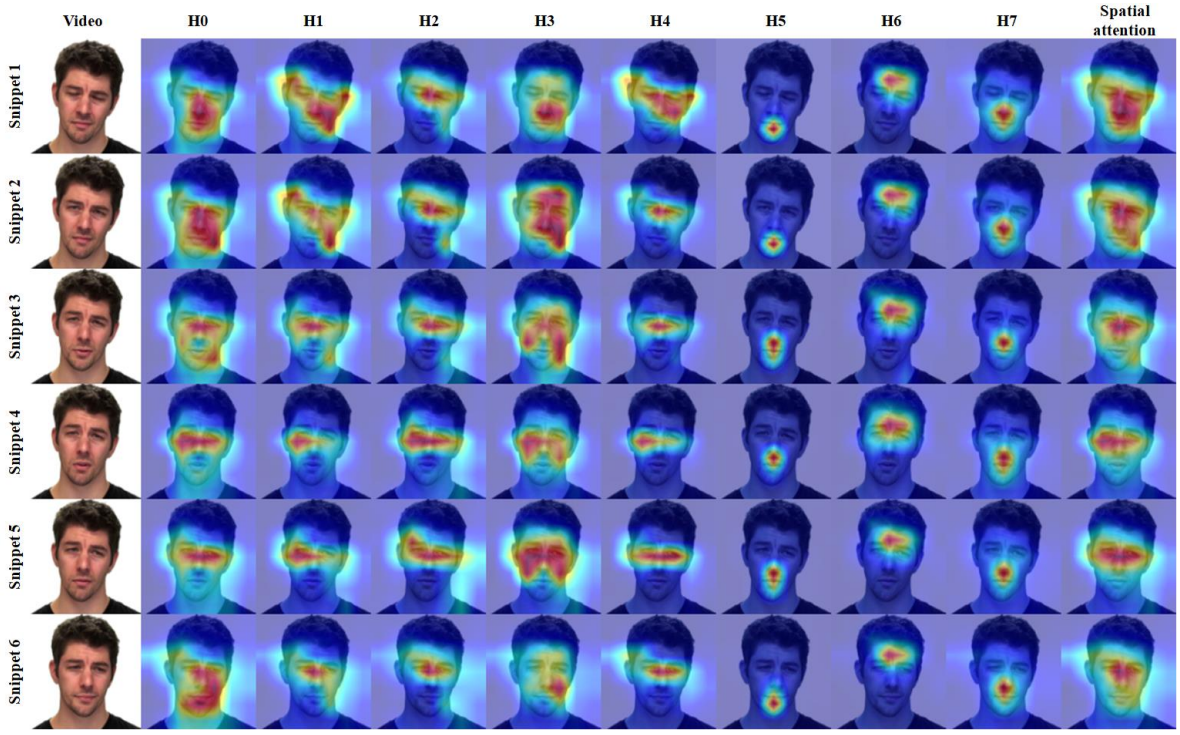

Fig. 8. Visualization of the learned spatial attention heads/regions (H_0 – H_7) for an actor depicting the sadness

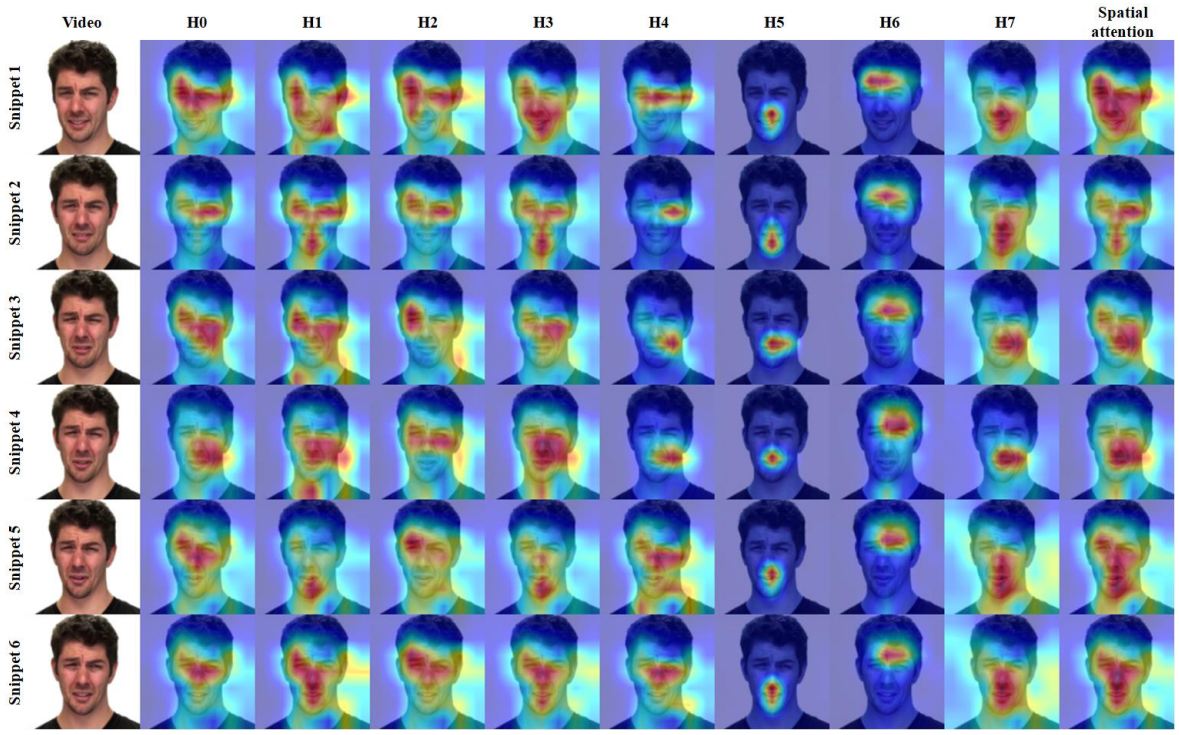

Fig. 9. Visualization of the learned spatial attention heads/regions (H_0 – H_7) for an actor depicting the disgust

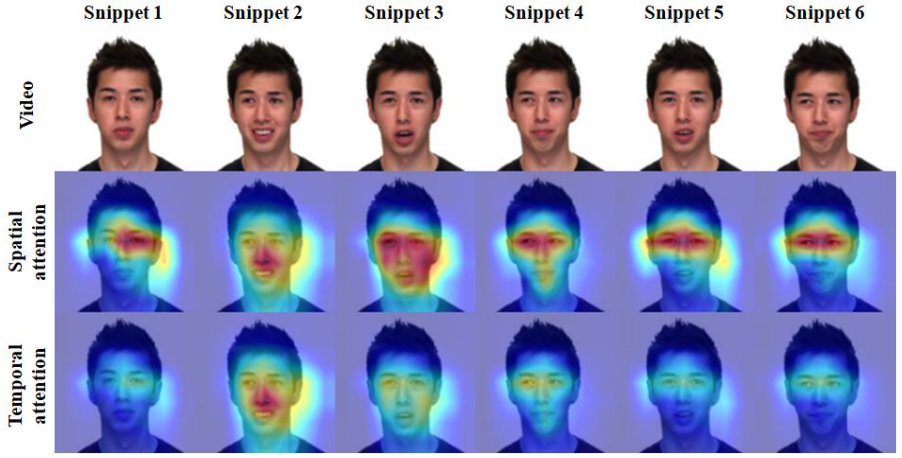

Fig. 10. Visualization of the learned spatial and temporal attention regions/video frames for an actor depicting the fear

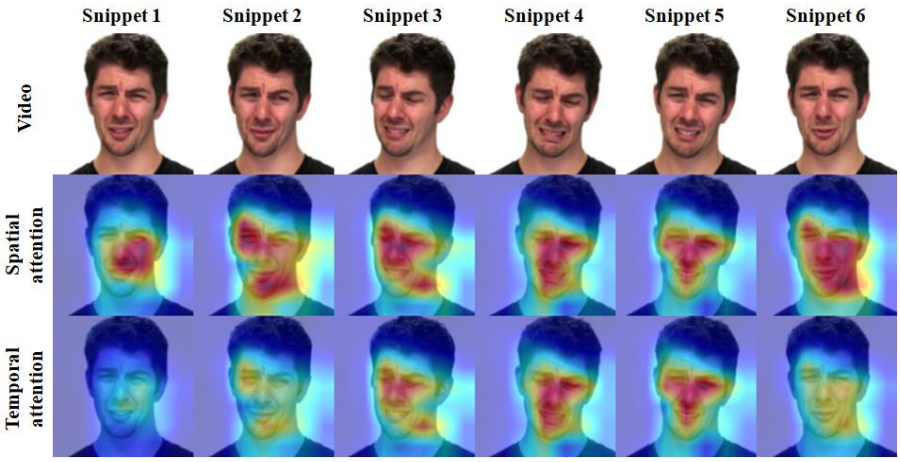

Fig. 11. Visualization of the learned spatial and temporal attention regions/video frames for an actor depicting the disgust

The following section presents the main drawbacks and limitations of the proposed framework architecture.

4.7. Limitations

The experimental evaluation conducted also makes it possible to identify some limitations of the proposed methodology:

(1). *The unavailability of comprehensive, labelled video dataset:* We have trained our framework on two publicly available datasets RAVDESS and CREMA-D showing actors of median age,between 20 and 50 years old, performing different discrete emotions. Thus, applying such architecture people with an age out of this range (e.g., older people) may reduce the accuracy scores because of the facial morphology variations involved. In addition, all emotions in the datasets are synthetically generated and may differ from natural expressions. Moreover, both databases are acquired in controlled environments. In real-life scenarios, the model may show lower performances because of the challenging lighting conditions or the partial occlusion of the subject's face.

(2). *Sensitivity to face pose variation:* A different drawback of the system relates to its inability to cope with important head pose variation (*i.e.,* superior to 20 degrees). A strategy to solve such an issue is to include into the process 3D facial data with depth information.

(3). *The inability of identifying micro-emotions:* The micro-emotions are brief expressions a person exhibits in high-stakes situations while showing their feelings. Such emotions may reveal the actual real state of a person but are very difficult to be identified by an automatic system or even by a novice person.

(4). *The capacity to detect only primary/basic emotions:* The proposed system has been designed to detect only the primary emotional states. However, a person may perform many secondary emotional states (such as: frustration, depression, satisfaction, etc). To deal with such cases additional training data with corresponding annotations is required.

5. Conclusions and perspectives

In this paper, we have proposed a novel deep attention model for discrete emotion recognition. The proposed approach is based on multimodal audio and visual information fusion and is designed to leverage the mutually complementary nature of features while maintaining the modality-specific information. From the methodological point of view, the core of the approach relies on: (1). an intra-modal attention mechanism that takes full advantage of the CNN characteristics to yield attentive (spatial, channel-wise and temporal) visual and audio features; (2). a cross-attention mechanism, that fusion the A-V data representations and efficiently combines the modalities in a complementary fashion; (3). a novel loss function that extends the triplet loss with a *polarity constraint* that takes into consideration the relations between the discrete emotion classes, designed to improve the latent space data

representation. By considering various attention mechanisms our model can better focus on discriminative face regions and relevant keyframes. In addition, the emotion constraint can guide the attention generation.

The experimental results conducted on two popular benchmarks RAVDESS [4] and CREMA-D [5] validate the proposed framework which achieves state-of-the-art performances with average accuracy scores of 89.25% and 84.57%, respectively. In addition, when compared to other methods [37], [56], [57], [27], [41], [58], [59] our methodology demonstrates its superiority with gains in accuracy ranging in the [1.72%, 11.25%] interval.

For future work, we intend to extend the architecture to include both emotion classification and regression tasks. Furthermore, the performance of the model can be increased by taking into consideration the information included in the textual channel (*i.e.*, speech to text transcripts, subtitle/close caption documents).

Acknowledgements

This work has been partially supported by a grant of the Romanian Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-TE-2021-0393, within PNCDI III.

References

[1]. Venkataramanan, K.; Rajamohan, H.R. Emotion Recognition from Speech. 2019. *arXiv:* 1912.10458.
[2]. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **1971**, 17, 124–129.
[3]. Ekman, P. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychol. Bull.* ***1994***, 115, 268–287.
[4]. Livingstone, S.R.; Russo, F.A. 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* ***2018***, 13.
[5]. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Trans. Affect. Comput.* **2014**, 5, 377–390.
[6]. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the *18th ACM International Conference on Multimedia*, Firenze, Italy, 25–29 October 2010; Association for Computing Machinery: New York, NY, USA, **2010**; pp. 1459–1462.
[7]. Boersma, P.; Weenink, D. PRAAT, a system for doing phonetics by computer. *Glot Int.* **2001**, 5, 341–345.
[8]. Bhavan, A.; Chauhan, P.; Hitkul; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst*. **2019**, 184, 104886.
[9]. Mao, Q.; Dong, M.; Huang Z.; Zhan, Y.; Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks, in IEEE Transactions on Multimedia, **2014**, vol. 16, no. 8, pp. 2203-2213.
[10]. Pepino, L.; Riera, P.; Ferrer, L. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September **2021**; pp. 3400–3404.
[11]. Ma, X.; Wu, Z.; Jia, J.; et al, Speech Emotion Recognition with Emotion-Pair based Framework Considering Emotion Distribution Information in Dimensional Emotion Space, Proc. Interspeech **2017**, pp.1238-1242, 2017.
[12]. Lian, Z.; Li, Y.; Tao, J.; Huang, J.; Speech emotion recognition via contrastive loss under siamese networks, Proc. of ASMMC-MMAC, pp. 21-26, **2018**.
[13]. Issa, D.; Demirci, M. F.; Yazici, A.; Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control, **2020**, vol. 59, p. 101894.
[14]. Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Learning Utterance-Level Representations with Label Smoothing for Speech Emotion Recognition, *Proc. Interspeech*, **2020**, 4079-4083.
[15]. Mocanu, B.; Tapu, R.; Zaharia, T. Utterance Level Feature Aggregation with Deep Metric Learning for Speech Emotion Recognition. *Sensors* **2021**, *21*, 4233.
[16]. Mirsamadi, S.; Barsoum, E.; Zhang, C.; Automatic speech emotion recognition using recurrent neural networks with local attention, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), **2017**, pp. 2227-2231.
[17]. Tzinis, E.; Potamianos, A.; Segment-based speech emotion recognition using recurrent neural networks, 2017 *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, **2017**, pp. 190-195.
[18]. Huang, J.; Li, Y.; Tao, J.; Speech emotion recognition from variable-length inputs with triplet loss function, *In Interspeech*, **2018**, pp. 3673–3677.
[19]. Atila, O.; Sengür, A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Appl. Acoust.*, **2021**, 182, 108260.
[20]. Mustaqeem; Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.*, **2021**, 102, 107101.
[21]. Wijayasingha, L.; Stankovic, J.A. Robustness to noise for speech emotion classification using CNNs and attention mechanisms., *Smart Health* **2021**, 19, 100165.

[22]. Nguyen, B.T.; Trinh, M.H.; Phan, T.V.; Nguyen, H.D. An efficient real-time emotion detection using camera and facial landmarks. *In Proceedings of the 2017 Seventh International Conference on Information Science and Technology (ICIST)*, Da Nang, Vietnam, 16–19 April **2017**; pp. 251–255.

[23]. Bagheri, E.; Esteban, P.G.; Cao, H.L.; De Beir, A.; Lefeber, D.; Vanderborght, B. An Autonomous Cognitive Empathy Model Responsive to Users' Facial Emotion Expressions. *Acm Trans. Interact. Intell. Syst*. **2020**, 10, 20.

[24]. Tautkute, I.; Trzcinski, T. Classifying and Visualizing Emotions with Emotional DAN. *Fundam. Inform.* **2019**, 168, 269–285.

[25]. Jaderberg, M.; Simonyan, K.; Zisserman, A.; kavukcuoglu, K. Spatial Transformer Networks. *In Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, **2015**; Volume 28.

[26]. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* **2021**, 21, 3046.

[27]. Luna-Jiménez, C.; Cristóbal-Martín, J.; Kleinlein, R.; Gil-Martín, M.; Moya, J.M.; Fernández-Martínez, F. Guided Spatial Transformers for Facial Expression Recognition. *Appl. Sci.* **2021**, 11, 7217

[28]. D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 *IEEE International Conference on Computer Vision (ICCV)*, **2015**, pp. 4489-4497.

[29]. I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3d convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **2017**, pp. 1609–1618.

[30]. Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM*, **2016**, pp. 445–450.

[31]. X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, **2017**, pp. 577–582.

[32]. J. Zhao, X. Mao, and J. Zhang, "Learning deep facial expression features from image and optical flow sequences using 3d cnn," *The Visual Computer*, pp. 1–15, **2018**.

[33]. D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, *Comput. Vis. Image Underst*. 174 (**2018**) 33–42.

[34]. H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang, S. Feng, Chinese multimodal emotion recognition in deep and traditional machine learning approaches, in: 2018 *First Asian Conference on Affective Computing and Intelligent Interaction* (ACII Asia), IEEE, **2018**, pp. 1–6.

[35]. S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al. *Combining modality specific deep neural networks for emotion recognition in video*, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, **2013**, pp. 543–550.

[36]. Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., & Keutzer, K. An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, **2020**, *34*(01).

[37]. E. Ghaleb, J. Niehues and S. Asteriadis, "Multimodal Attention-Mechanism For Temporal Emotion Recognition," *2020 IEEE International Conference on Image Processing (ICIP)*, **2020**, pp. 251-255.

[38]. Wang, Y., Wu, J., Heracleous, P., Wada, S., Kimura, R., & Kurihara, S. (2020). Implicit Knowledge Injectable Cross Attention Audiovisual Model for Group Emotion Recognition. In *ICMI* **2020** - *Proceedings of the 2020 International Conference on Multimodal Interaction* pp. 827-834.

[39]. S. Parthasarathy and S. Sundaram, "Detecting Expressions with Multimodal Transformers," *2021 IEEE Spoken Language Technology Workshop (SLT)*, **2021**, pp. 636-643.

[40]. P. Tzirakis, J. Chen, S.Zafeiriou, B. Schuller, "End-to-end multimodal affect recognition in real-world environments", *Information Fusion*, Volume 68, **2021**, Pages 46-53.

[41]. A. I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, *Knowledge-Based Systems*, Volume 244, **2022**.

[42]. Mikels, J. A.; Fredrickson, B.L.; Larkin, G.R. et al.; Emotional category data on images from the International Affective Picture System, *Behavior Res. Methods* 37(4):626–630, **2005.**

[43]. K. Hara, H. Kataoka and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2018**, pp. 6546-6555, doi: 10.1109/CVPR.2018.00685.

[44]. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2016**, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[45]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, I.Polosukhin, "Attention is all you need", *In Advances in Neural Information Processing Systems*, pp. 6000–6010, **2017**.

[46]. Nawab, S. H.; Quatieri, T. F., Short-time Fourier transform, in Advanced Topics in signal processing, J. S. Lim and A. V. Oppenheim, Eds.Upper Saddle River, NJ, USA: *Prentice-Hall*, **1987**, pp. 289–337.

[47]. Huang, J.; Li, Y.; Tao, J.; Lian, Z.; Wen, Z.; Yang, M.; Yi, J. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October **2017**; pp. 11–18.

[48]. S. Zhang, S. Zhang, T. Huang, W. Gao and Q. Tian, "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030-3043, Oct. **2018**, doi: 10.1109/TCSVT.2017.2719043.

[49]. B. T. Atmaja and M. Akagi, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **2020**, pp. 4482-4486, doi: 10.1109/ICASSP40776.2020.9052916.

[50]. J. Liu *et al*., "Multimodal Emotion Recognition with Capsule Graph Convolutional Based Representation Fusion," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **2021**, pp. 6339-6343, doi: 10.1109/ICASSP39728.2021.9413608.

[51]. L. Sun, B. Liu, J. Tao and Z. Lian, "Multimodal Cross- and Self-Attention Network for Speech Emotion Recognition," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, **2021**, pp. 4275-4279, doi: 10.1109/ICASSP39728.2021.9414654.

[52]. J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, **2009**, pp. 248-255.

[53]. J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2017**, pp. 4724-4733.

[54]. D. P. Kingma and M. Welling."Auto-encoding variational bayes". ICLR, **2014.**

[55]. X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *ACAIS*, volume 9, pages 249–256, **2010**.

[56]. L. Su, C. Hu, G. Li, D. Cao, "MSAF: Multimodal Split Attention Fusion", *arXiv preprint* arXiv: 2012.07175, **2020**.

[57]. Z. Fu, F. Liu, H. Wang, J. Qi, X.Fu, A. Zhou, Z. Li, "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition", *arXiv preprint* arXiv: 2012.07175, **2021**.

[58]. L. Goncalves and C. Busso, "AuxFormer: Robust Approach to Audiovisual Emotion Recognition," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7357-7361.

[59]. Chang, X.; Skarbek, W. Multi-Modal Residual Perceptron Network for Audio–Video Emotion Recognition. *Sensors* **2021**, *21*, 5452.

[60]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *2017 IEEE International Conference on Computer Vision (ICCV)*, **2017**, pp. 618-626.

[22]. Y. Li, C. Papayiannis, V. Rozgic, E. Shriberg and C. Wang, "Confidence Estimation for Speech Emotion Recognition Based on the Relationship Between Emotion Categories and Primitives," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 7352-7356

[23]. A. Ghriss, B. Yang, V. Rozgic, E. Shriberg and C. Wang, "Sentiment-Aware Automatic Speech Recognition Pre-Training for Enhanced Speech Emotion Recognition," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 7347-7351.

[24]. S. Sahu, R. Gupta, G. Sivaraman and C. Espy-Wilson, "Smoothing Model Predictions Using Adversarial Training Procedures for Speech Based Emotion Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 4934-4938.

[25]. Z. Ren, A. Baird, J. Han, Z. Zhang and B. Schuller, "Generating and Protecting Against Adversarial Attacks for Deep Speech-Based Emotion Recognition Models," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 7184-7188.

[26]. Su, B.-H., Lee, C.-C. "Vaccinating SER to Neutralize Adversarial Attacks with Self-Supervised Augmentation Strategy", *Proc. Interspeech 2022*, 1153-1157.

[27]. Parry, J., DeMattos, E., Klementiev, A., Ind, A., Morse-Kopp, D., Clarke, G., Palaz, D. "Speech Emotion Recognition in the Wild using Multi-task and Adversarial Learning", *Proc. Interspeech 2022*, 1158-1162.

[28]. Gudmalwar, A., Basel, B., Dutta, A., Rama Rao, C.V. "The Magnitude and Phase based Speech Representation Learning using Autoencoder for Classifying Speech Emotions using Deep Canonical Correlation Analysis", *Proc. Interspeech 2022*, 1163-1167.

[33]. A. V. Savchenko, L. V. Savchenko and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," in *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132-2143, 1 Oct.-Dec. 2022.

[34]. M. Pourmirzaei, G. A. Montazer, and F. Esmaili, "Using self-supervised auxiliary tasks to improve fine-grained facial representation," 2022, *arXiv:2105.06421*.

[35]. F. Ma, B. Sun and S. Li, "Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion," in *IEEE Transactions on Affective Computing*, 2022.

[36]. F. Xue, Q. Wang and G. Guo, "TransFER: Learning Relation-aware Facial Expression Representations with Transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 3581-3590.

[37]. P. Antoniadis, P. P. Filntisis and P. Maragos, "Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition," *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Jodhpur, India, 2021, pp. 1-8.

[38]. N. I. Abbasi, S. Song and H. Gunes, "Statistical, Spectral and Graph Representations for Video-Based Facial Expression Recognition in Children," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 1725-1729.

[39]. R. Miyoshi, S. Akizuki, K. Tobitani, N. Nagata and M. Hashimoto, "Convolutional Neural Tree for Video-Based Facial Expression Recognition Embedding Emotion Wheel as Inductive Bias," *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022, pp. 3261-3265

[42]. D. Hu, X. Hou, L. Wei, L. Jiang and Y. Mo, "MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 7037-7041.

[43]. J. Zhao, R. Li, Q. Jin, X. Wang and H. Li, "Memobert: Pre-Training Model with Prompt-Based Learning for Multimodal Emotion Recognition," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 4703-4707.

[44]. J. Zhao, G. Ru, Y. Yu, Y. Wu, D. Li and W. Li, "Multimodal Music Emotion Recognition with Hierarchical Cross-Modal Attention Network," *2022 IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, 2022, pp. 1-6.

[45]. H. -D. Le, G. -S. Lee, S. -H. Kim, S. Kim and H. -J. Yang, "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning," in *IEEE Access*, vol. 11, pp. 14742-14751, 2023.

[46]. W. Chen, X. Xing, X. Xu, J. Yang and J. Pang, "Key-Sparse Transformer for Multimodal Speech Emotion Recognition," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 6897-6901.

[47]. V. John and Y. Kawanishi, "Audio and Video-based Emotion Recognition using Multimodal Transformers," *2022 26th International Conference on Pattern Recognition (ICPR)*, Montreal, QC, Canada, 2022, pp. 2582-2588.

[61]. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[62]. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv*:1705.06950, 2017.

[63]. Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R Manmatha, M. Li. "A comprehensive study of deep video action recognition". *arXiv preprint arXiv*:2012.06567, 2020.

[64]. W.B. Cannon, "The James-Lange theory of emotions: A critical examination and an alternative theory." *Am. J. Psychol*. 1927, 39, 106–124.

[65]. Shu, L.; Xie, J.; Yang, M.; Li, Z.; Li, Z.; Liao, D.; Xu, X.; Yang, X. A Review of Emotion Recognition Using Physiological Signals. *Sensors* **2018**, *18*, 2074.