# Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition

Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis
Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, the Netherlands
{esam.ghaleb, mirela.popa, stelios.asteriadis}@maastrichtuniversity.nl

***Abstract*—In Audio-Video Emotion Recognition (AVER), the idea is to have a human-level understanding of emotions from video clips. There is a need to bring these two modalities into a unified framework, to effectively learn multimodal fusion for AVER. In addition, literature studies lack in-depth analysis and utilization of how emotions vary as a function of time. Psychological and neurological studies show that negative and positive emotions are not recognized at the same speed. In this paper, we propose a novel multimodal temporal deep network framework that embeds video clips using their audio-visual content, onto a metric space, where their gap is reduced and their complementary and supplementary information is explored. We address two research questions, (1) how audio-visual cues contribute to emotion recognition and (2) how temporal information impacts the recognition rate and speed of emotions. The proposed method is evaluated on two datasets, CREMA-D and RAVDESS. The study findings are promising, achieving the state-of-the-art performance on both datasets, and showing a significant impact of multimodal and temporal emotion perception.**

***Index Terms*—multimodal and incremental learning, deep metric learning, audio-video emotion recognition**

## I. Introduction

Emotions are a key component in human-human communications, and a great amount of affective information is displayed through facial expressions, gestures, speech, and other means [1]–[3]. However, HCI still lacks these elements to enable a more human-centered interaction. Human-centered computation through affective computing can help in recognizing emotions, and generate proper actions to have richer communication. Applications of affective computing range from education [4], autonomous driving [5], entertainment [6] to health-care [7].

From a psychological and neurological perspective, research has gained preliminary evidence about how the brain tends to bind information received from different modalities [8]. The binding interaction of different modalities has been demonstrated in the so-called McGurk effect [9], which shows that audio signals can be altered by the display of incongruent visual information. This multimodal integration is essential for multimodal perception in many cases [10], since it enables accurate perception in a noisy environment or in a state of confusion. In addition, studies show that speech and facial expressions can substitute and complement each other in many tasks such as emotion recognition or speech identifications [8].

Another question of multimodal emotion perception is incremental perception. Emotion recognition varies as a function of time, and this temporal process might change according to the kind of emotions [8]. However, much of the focus in AVER systems relies on multimodal learning and fusion, through the selection of apex moments [11], and there is still less exploration of multimodal interaction over time [12]. In the literature of multimodal emotion recognition, much of the effort went either to a late fusion of modalities [13] or to building temporal features, based on the assumption that emotions are expressed simultaneously and global information can be obtained to represent the emotional content of a video clip. However, these studies overlooked an important aspect of how multimodal information binds and evolves over time, which is the aim of this research paper.

In this work, we address the following research questions: 1. how to efficiently connect information from different modalities, and 2. how to deal with incremental emotion display. For this purpose, we design a data-driven unified multimodal-temporal deep learning methodology to explore the variation of emotion expression over time through audio-visual modalities. The proposed method aligns the visual and audio representations using multi-stages integration and learning. The integrated multimodal framework is inspired by a gating paradigm introduced in [14] by F. Grosjean. In this paradigm, a stimulus is presented in successive segments of increasing duration. It is shown that emotion perception improves over time, by providing people with a richer context.

Furthermore, by employing an efficient metric distance, the accuracy of many classification and retrieval problems [15], [16] can be increased, as it contributes to obtaining an improved performance and robust representation. In metric learning, the task is to learn a distance function that is efficient to measure the similarity and dissimilarity of data samples. It is a successful technique in many domains, such as person identification and image classification and retrieval [17].

In this paper, we make two main contributions. First, we propose a novel end-to-end multimodal deep metric learning architecture. Our integrated temporal paradigm aims to learn audio-visual embeddings (representations) that are aware of emotional content in both auditory signals and facial expressions. The proposed methodology is illustrated in Fig. 1 and explained in details in Section III. Second, we develop an effi-

978-1-7281-3888-6/19/$31.00 

cient learning algorithm through a multimodal and incremental triplet sets' mining and data augmentation, which is crucial to train the proposed methodology and improve the performance. The proposed solutions are explained in Section IV. Finally, we demonstrate these contributions in Section V, by providing an extensive evaluation on two large-scale audio-video emotion datasets, namely: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18] and Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [19], showing how the proposed method efficiently fuses the different contributions of each modality over time.

## II. Related Work

**Multimodal emotion recognition:** human emotion recognition is a multi-disciplinary domain that has gained notable attention in the fields of AI, HCI, psychology, and related fields [3]. Humans rely on diverse modalities in their social interactions such as facial expressions and speech [8]. Similarly, multimodal learning in affective computing seems natural and has been proven to increase system robustness and accuracy [1]. In addition, with the recent developments in Deep Learning (DL) architectures [20], deep learning has been applied in AVER using shallow and deep networks.

**Temporal emotion recognition:** emotion perception might vary over time according to classes of emotions. Previous studies on AVER indicate the importance of multimodal and temporal information [1]. However, the existing systems on emotion recognition focus less on studying how to bind multimodal information over time. Many studies attempt to model onset, apex and offset of expressions. In addition, some studies select apex expressions and speaking face tracks for multimodal learning and fusion [11]. In [12], Yelin Kim and Emily M. Provost proposed a framework to explore timing and duration's effect on emotion classes. Their work is based on window averaging of audio-visual cues to spot the influential regions over time for utterance-level emotion inference.

Nevertheless, our approach focuses on building multimodal incremental embeddings and checking how they contribute to emotion recognition over time. The proposed paradigm benefits from initial time windows of emotion expression and transmits this knowledge to the subsequent windows.

**Multimodal learning:** multimodal learning is one of the challenging frontiers in machine learning [2]. Different approaches have been proposed for multimodal learning and can be categorized as follow [21]: multi-representation alignment (e.g. correlation-based models [22] and distance and similarity-based models [23]) and multi-view representation fusion (e.g. graphical models [21] and neural network models [24]).

Our work follows the category of multi-view representation alignment, by employing Deep Metric Learning (DML). The basic concept of metric learning is to modify a conventional metric, such as Euclidean distance, by including an efficient mapping function: $f : x \rightarrow R^n$. In this mapping process, the aim is to bring similar samples closer, and the dissimilar ones further, given the distance: $d(f(x_i), f(x_j) = \|f(x_i) - f(x_j)\|_2^2$. Conventional metric learning approaches, such as Large Margin Nearest Neighbours (LMNN), usually learn a linear mapping. This linear mapping might suffer from the non-linear relationships between data samples, especially in multimodal learning tasks.

In our study, we adopt DML given its efficiency in learning robust representations. Involving distance metric embedding within the proposed architecture guarantees learning compact and discriminative features [25], [26]. Moreover, we employ triplet loss, since it involves negative and positive samples for a given anchor, which makes it more suitable for a classification task such as the one in AVER (more details are given in Section III-B). Another reason is that DML can tackle the lack of sufficient data to train deep models, since it exploits the data similarity between samples, which generates a larger pool of data to train DL efficiently.

**DML:** DL has been widely accepted as an effective model in highly non-linear data, and currently is proven to be the state-of-the-art in data representation and perception tasks [2], [20]. Therefore, DL can be used explicitly in learning mapping functions for metric learning, through a set of non-linear transformations. Moreover, multimodal deep learning is a way to exploit the dependencies and complementary information in multimodal tasks such as AVER [2], [27].

## III. Methodology

In this paper, we aim to generate temporal audio-visual embeddings for accurate multimodal and temporal (incremental) emotion perception. Specifically, we aim at producing discriminative embeddings, by taking into consideration the binding information between audio-visual modalities diachronically. When designing the multimodal deep learning framework based on metric loss, we have these objectives:

- It should exploit the complementary information in the audio-visual representations.
- It is expected to produce discriminative and representative features and to reduce the gap between the audio-visual representations. In AVER, one of the challenges is that usually there is not a perfect alignment between the two data channels in terms of emotion expression. For example, happiness could be initially expressed through facial expressions, while corresponding time-slices in the audio channels are not useful yet. However, the following audio time slices could provide valuable information [28].
- The framework should take into consideration the temporal evolution of emotion expression in video-clips. Emotion expression could have gradual descent or ascent, where emotion expression is peaked at certain moments. Studies demonstrated that emotion perception might require a different amount of time for an accurate detection [12]. Therefore, these alterations could be exploited efficiently through a temporally-trimmed framework.

### A. Temporal Joint Embeddings

Fig. 1 outlines the proposed architecture for achieving incremental shared representations, modeling the relationships

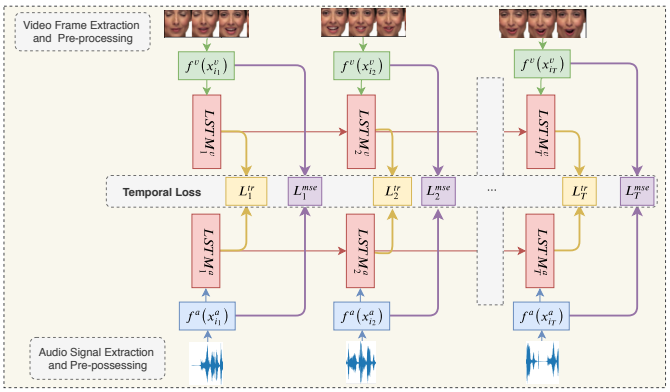



Fig. 1: The proposed framework of multimodal temporal deep metric learning for AVER. It has two streams of audio ($f^a(x^a)$) and video ($f^v(x^v)$) sub-networks, diachronically connected via LSTM cells as a gating paradigm. In each gate, identification and discriminative signals guide the training of the network.

between the two modalities over time. This is pursued through identification and discriminative signals in each time-window that are averaged to obtain a temporal framework. The proposed architecture shows how the sub-networks are connected incrementally via LSTM cells.

To target our research objectives, we learn joint embeddings via two stream networks, audio, and visual networks. In each gate, we employ 3D-CNN [29] for visual mapping: $f^v(x^v) : \mathbb{R}^{d^v} \rightarrow \mathbb{R}^p$ and soundNet [30] for audio mapping: $f^a(x^a) : \mathbb{R}^{d^a} \rightarrow \mathbb{R}^p$, which map the visual and audio cues of a video clip onto a shared space: $R^p$. In each sub-network, gates are connected by Long-Short-Term-Memory (LSTM) cells to explore and utilize temporal dependency between video clip segments (time-windows). Prior to feeding LSTM cells with the visual mapping, the gap between the two modalities is reduced, through minimizing the distance between the two representations via Mean Square Error (MSE) loss functions.

### B. Definitions

In AVER, a dataset ($\mathbb{D}$) contains $m$ video clips with audio and visual signals, while each clip is annotated with a discrete emotion $I^y$:

$$\mathbb{D} = \{(x_1^v, x_1^a, I^{y^1}), (x_2^v, x_2^a, I^{y^2}), ..., (x_m^v, x_m^a, I^{y^m})\}$$

$x_i^v \in X^{d^v}$ and $x_i^a \in X^{d^a}$ denote video and audio feature vectors corresponding to the $i^{\text{th}}$ sample of video $X^{d^v \times m}$ and audio $X^{d^a \times m}$ data samples. $I^{y^j} \in I^Y$ refers to the given discrete emotion label. The goal, in AVER, is to predict the emotional content of a given sample test.

In our work, we apply DML based on triplet networks. The loss function of this type of architecture uses triplet sets: $(x_i, x_i^+, x_i^-)$, where $x_i$ is an anchor, $x_i^+$ and $x_i^-$ are similar and dissimilar examples to $x_i$, respectively (as shown in (1)). The optimization procedure aims to minimize the distance between the anchor (baseline) input to a positive pair while maximizing the distance from the anchor to the negative pair [26].

$$d_f(x_i, x_i^+, x_i^-) = \|f(x_i) - f(x_i^+)\|_2^2 - \|f(x_i) - f(x_i^-)\|_2^2 + \text{margin} \quad (1)$$

### C. Formulation

To learn the parameters of each audio-visual mapping, and the temporal connections between the cells, taken into consideration our research objectives, in each time-window (gate), we employ a temporal metric that has two terms: (1) multimodal triplet $L^{tr}$ and (2) MSE $L^{mse}$ loss functions:

$$\begin{aligned} \underset{f^v, f^a}{\text{argmin}}\ \mathrm{L} &= \frac{1}{2T}\sum_{t=0}^{T} L_t = L_t^{mse} + L_t^{tr} = \frac{1}{N}\sum_{i=0}^{N} \|f^v(x_{it}^v) - f^a(x_{it}^a)\|_2^2 \\ &+ \frac{1}{2N}\sum_{i=0}^{N}\sum_{m=v,a} max\big(d_{f^m}(x_{it}^m, x_{it}^{+m}, x_{it}^{-m}), 0\big) \end{aligned} \quad (2)$$

where $t$ refers to a time-window in the architecture, which can be up to T; N is the number of samples per mini-batch; $x_{it}^{a,v}$ indicate the corresponding segment of ($a$) audio or ($v$) video data in a given ($x_i$) video clip. We formulate the multimodal triplet loss $L^{tr}$ that optimizes $f^v(x^v)$ and $f^a(x^a)$ to minimize the distance between an anchor and a positive sample, while increasing the distance to a negative sample.

In each time-window and mini-batch, for both modalities, we sample a two sets $T_{a,v}$ of triplets: $(x_{it}^{a,v}, x_{it+}^{a,v}, x_{it-}^{a,v})$, where $x_{it}^{a,v}$ is an anchor, and $x_{it+}^{a,v}$ and $x_{it-}^{a,v}$ are similar and dissimilar examples to $x_{it}^{a,v}$, respectively. In other words, triplet loss minimizes the intra-class variations and maximizes the inter-class variations and provides an identification signal, which is conducted on the audio and video embeddings through a multimodal and incremental negative and positive triplet sets mining (explained in Sect. III-D).

The second term of the temporal loss formulation, $L^{mse}$, is responsible for leveraging similar information in both modalities, by minimizing the distance between the audio and video embeddings. Therefore, the main advantage of our framework consists in not only capturing the complementary and supplementary information between the audio-video channels in a global manner, but also in modeling them across time, contributing to a better overall emotion understanding, regarding its display pattern.

### D. Multi Windows Triplet Sets Mining (MWTSM)

One of the main challenges in DML is that triplet loss often suffers from slow convergence. In each time-window, as the possible number of the triplet sets is large, DML learns to map correctly easy samples. However, hard negative mining is essential to improve the performance of the network and to provide it with useful training guidance. Therefore, as suggested by [25], [31], online hard negative mining based on mini-batches is an effective solution.

In our work, we propose MWTSM, an effective technique for hard samples mining, at each time-window (t), where we opt for hard negative mining by harvesting the triplets based on temporal-multimodal embeddings. Each mini-batch contains N-samples (video-clips) for each modality. Therefore, for each sample in each modality, we obtain triplet-sets with hard negative samples that have the largest distances according to the defined metric in Eq. 1, resulting in a total of $2N$ triplet sets. Considering that in our framework we have $T$

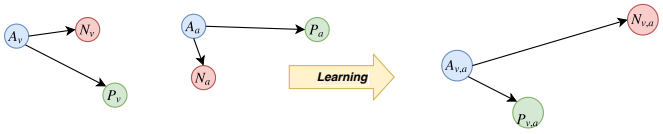



Fig. 2: Example of hard negative triplet sets mining based on audio-video modalities. The proposed DML minimizes the distance between the anchor video-clip and its positive sample, while maximizing the distance to its negative sample.

time-windows, the final number of triplet sets is $2TN$. Fig 2 displays this process to a given video-clip for two modalities.

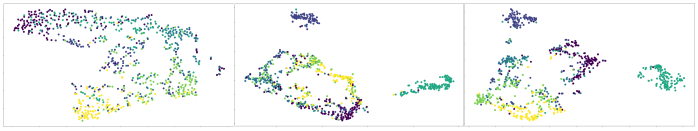
In addition, in each time-window (gate), the hard sets of triplets are forwarded to the next windows. Such that, in the new time-window, the hard negative mining selects new samples for each anchor. This strategy enables the framework to have many useful triplets and avoids repeated sets to optimize the learning process.

## IV. Implementation Details

**Data augmentation in DML for AVER:** during our implementation and experimentation, we observed that random data augmentation for metric learning in affective computing could be quite harmful, degrading both the learning process and the performance. The reason is that embeddings of different facial regions could trigger a false selection of triplet sets. Therefore, for each mini-batch, we apply similar methods for data augmentation and they can be random for different mini-batches, but not within one of them. For example, cropping is applied by re-sizing video's frames into $112 \times 112$ resolutions, then randomly cropping $96 \times 96$ patches. Similarly, data augmentation techniques are applied on audio raw signals such as adding noise, changing the pitch, and the speed of the signal. However, during the test, models are applied over given video clips segments and, for visual mapping, we select $96 \times 96$ center crops. In addition, data augmentation is not applied to audio signals, during the testing and evaluation phases.

**Audio and visual mapping:** in our study, for the audio and video mapping, we employed SoundNet [30] and the 3D-CNN branch of Two-Stream Inflated 3D ConvNet (I3D) [29], respectively. More details about the architecture of these models are given in the corresponding papers. The two architectures are pre-trained using Softmax approaches. In our work, we adopted these architectures as audio and visual mapping models for the proposed framework, due to the following reasons: (1) their ability to capture and model temporal data up to several seconds, and (2) they are state-of-the-art models and have a good discriminative power for many other audio-video recognition tasks.

**Training procedure:** the proposed architecture is trained on two Titan XP GPUs for 100 epochs. The batch size is 5 times the number of emotion classes. We used Adam optimizer with a 0.001 learning rate, and 0.0005 weight decay. The margin in the triplet loss was set as 1. We found these parameters to give a good performance with a combination of other parameters such as the number of windows and their length (discussed in the experimental Section V-B).



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

Fig. 3: t-SNE plot for a subset of CREMA-D, in the learned subspace. In (a), (b) and (c), we visualize the audio, video and the concatenated audio-video data following the proposed methodology, where the clusters are better structured, and best separated when both modalities exist.

## V. Experiments

**Experimental Setup:** The multimodal perception is based on the Gating Paradigm (**GP**) proposed in [14]. In the standard GP, audio-visual cues are presented in successive segments, with a forward manner. Next, in each gate, participants give confidence to each segment. Similarly in our approach, during training and evaluation, these segments were presented to the framework, where the temporal layer, based on LSTM, is accumulating and learning the contribution of them. Our approach resembles the standard GP, in which participants rate segments, by evaluating the multimodal presentation in each step using the identification signal provided by the triplet loss.

Each audio-visual segment's length was set to either 1 or 2 seconds. As a result, each LSTM output corresponds to the number of the previous gates. For example, the first cell of LSTM has information of the first level (e.g. 1 second), the second (2 seconds), the third (3 seconds) and so on. These windows can have an overlap of 500 ms. Next, in each gate, audio-visual embeddings are assessed through the identification loss (triplet loss), and the gap between the two modalities is reduced by the MSE loss (as defined in 2).

### A. Evaluation Scenarios and Datasets

**Validation protocol:** the efficiency of the proposed methodology is evaluated on two public AVER datasets, CREMA-D [19] and RAVDESS [18]. We divided the two datasets into 10-folds to perform cross-validation based on subjects. In other words, for each fold, subjects' clips are either in the training or testing sets, and there is not any overlap between these two sets. Subsequently, in each fold, we trained the proposed framework on the training folds and then test it on the remaining fold. The reported results are the average of these 10-folds.

**Classification and evaluation scenarios:** in each time-window, LSTM output is considered the corresponding embeddings of the given gate. The dimension of these embeddings is 400. The embeddings produced by the proposed framework are suitable to be evaluated by a simple classifier such as K-Nearest Neighbor (KNN). The evaluation is applied on the audio, video and the concatenated audio-video embeddings. K was set to 15, while the distance used is the Euclidean distance. The method is tested according to the following scenarios:

- A baseline constructed via unimodal and multimodal perception based on global information (apex moments of the

TABLE I: Tests for Various Configurations. RAVDESS and CREMA-D have an average of $3.82 \pm 0.34$, and $2.63 \pm 0.53$ seconds length video clips, respectively.

| Dataset | #Windows | Length | Overlap | Accuracy % |
|---|---|---|---|---|
| RAVDESS | 2 | 1 | ✓ | 64.1 |
| | 2 | 2 | ✓ | 64.7 |
| | 4 | 1 | ✓ | 66.2 |
| | 6 | 1 | ✓ | 67.5 |
| | 8 | 1 | ✓ | **67.7** |
| | 2 | 1 | ✗ | 62.3 |
| | 4 | 1 | ✗ | 65.3 |
| CREMAD | 2 | 1 | ✓ | 73.7 |
| | 4 | 1 | ✓ | 73.5 |
| | 6 | 1 | ✓ | **74.0** |

video-clip) without the temporal loss (referred to as global approach).

- A baseline based on the concatenation of LSTM embeddings, trained using DML for both modalities. We refer to it as LSTM's feature concatenation.
- Incremental perception, based on the gating paradigm through the proposed method (our approach).

**CREMA-D** [19] is a large-scale multimodal emotion expression dataset. It contains 7442 clips from 91 actors (43 females and 48 males). Their age ranges between 20 and 74 and they come from a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic). Actors were asked to speak 12 sentences, according to six different emotions, anger, disgust, fear, happiness, neutral, and sadness, with four different levels (intensities), low, medium, high and unspecified.

**RAVDESS** [18] is a multimodal emotional speech and songs database. In this work, we chose to use the speech part of the dataset as it is labeled with eight archetypal emotions: anger, happiness, disgust, fear, surprise, sadness, calmness and neutral. This subset contains a total of 2880 recordings.

A **visualization** of embeddings from the final time-window is provided in Fig. 3. The figure illustrates the clusters formed based on emotion classes. More importantly, we can observe that the clustering is improved when the two modalities are combined, compared to only visual or audio information. In addition, in our experiments, we observe that the contribution of visual information in the multimodal perception is greater than the audio information.

### B. Model's Hyper Parameters Evaluation

We evaluated the framework parameters, such as the number of cells (windows), length of audio and video inputs, and whether these inputs are overlapping or not. Table I shows the results on these parameters. Due to different lengths of video-clips in CREMA-D and RAVDESS, the number of windows was set differently. The results show that increasing the number of windows with overlapping segments helps significantly to increase the performance. Overlapping has less impact due to the length of the considered audio-visual signals, which is either 1 or 2 seconds. The best performance was obtained with overlapping one-second segments, and the maximum possible number of windows, namely 8, and 6, for RAVDESS and CREMA-D, respectively.

TABLE II: The recognition accuracy of unimodal and multimodal embeddings, with and without the temporal and multimodal DML.

| Embeddings | CREMA-D | RAVDESS |
|---|---|---|
| AO-Global | 56.4 | 50.1 |
| A-LSTM | 50.2 | 40.1 |
| AO-Gating Paradigm | 57.0 | 45.3 |
| VO-Global | 63.1 | 60.2 |
| VO-LSTM | 66.8 | 60.5 |
| VO-Gating Paradigm | 65.0 | 60.1 |
| AV-Global | 69.0 | 65.7 |
| Concatenation of A-LSTM and V-LSTM | 72.9 | 65.8 |
| AV-Gating Paradigm | **74.0** | **67.7** |
| Human Perception: AV | **63.6** | **80.0** |
| Dual Attention with LSTM: AV [32] | 65.0 | - |

### C. Impact of MWTSM

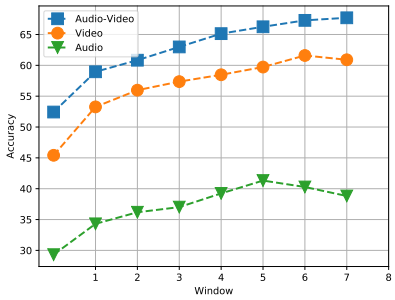
The strategy of MWTSM helps to increase the performance and guides the training procedure. This scheme is specifically important when the model starts to over-fit the training data. The anchors of triplet sets could have different options for positive and negative samples. Based on the hard negative mining, we selected the ones that are not previously chosen in the previous windows. In any given configuration, we noticed that the accuracy of the system increased by at least 3%. In addition, the proposed data augmentation helped the training in terms of generalization and learning process, unlike a total random augmentation that harmed the performance and prevented the system convergence.

### D. Uni-modal and Multimodal Evaluation

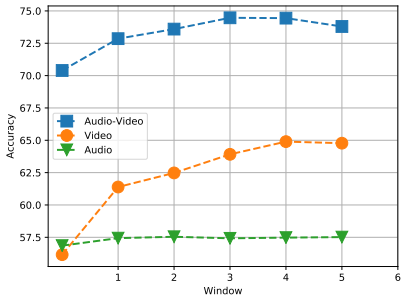
Fig. 4 gives a closer look into the recognition rates of the embeddings of audio-only (AO), video-only (VO), and audio-video (AV) modalities over time. Their representations are taken from the output of LSTM cells at each gate of the proposed framework. For both datasets, AV modality outperformed both AO and VO modalities. Most importantly, the results of multimodal perception prove that emotion recognition is a function of time, where rates are gradually ascending. Specifically, we notice that the rapid change of time is more obvious for the VO and AV modalities. This shows that our framework was able to utilize both the multimodal and the time impact for audio-video emotion recognition.

In addition, Table II illustrates the performance of the visual and audio embeddings obtained through our paradigm, and their concatenation as multimodal representations. We compare these performances to the embeddings of the visual and audio models which were trained using a global approach based on the apex moments of the video clips. We also provide the human-perception results reported in both datasets.

As shown in the table, perception rates increased when the two modalities embeddings are efficiently employed. In addition, comparing to the other two baselines, the gating paradigm increased significantly the accuracy by at least 2.9% and 2%, in CREMA-D and RAVDESS, respectively. Especially, with 74% accuracy, our method achieved better




(a) RAVDESS




(b) CREMA-D

Fig. 4: AV, VO, and AO accuracies over-time for RAVDESS and CREMAD.

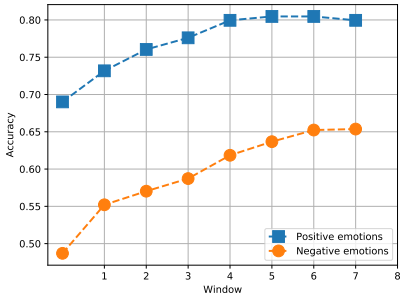
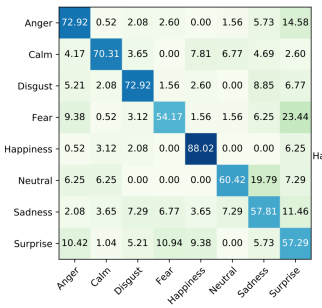
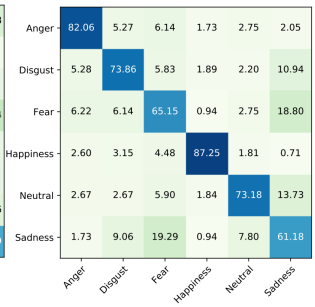



Fig. 5: Recognition speed of positive and negative emotions over-time.




(a) RAVDESS




(b) CREMA-D

Fig. 6: CM between true and predicted labels.

performance than the human-raters in CREMA-D and also better than the recently published results in [32]. In [32], the performance (65.0% accuracy) was obtained by combining facial and audio temporal features with LSTM. These results show the efficiency of our approach for enhanced joint multimodal learning and fusion. In addition, the proposed approach achieved significant results, with an accuracy of 67.7% on RAVDESS and also improved the recognition rates over the audio and video modalities. To the best of the authors' knowledge, these are the first results on the RAVDESS dataset, showing how the proposed framework captures dependencies and complementary information over time in AVER.

**Confusion Matrices (CMs):** finally, for each dataset, Fig. 6 displays the CMs between video-clips actual labels and the predicted labels in the last segment. The figure shows that the considered emotions are well discriminated, as the diagonal elements of the matrix have (by far) the highest accuracies. For example, in CREMA-D, the most common misleading happens between fear and sadness, with 19%. Most of these observations are aligned with the study presented in [19], where the predictions are based on human raters. In RAVDESS [18], human raters (with average recognition of 72.3) confused calm and neutral greatly, which is not the case in our system. Moreover, our framework eliminated most of the confusion between all the emotions and neutral state. According to the CM in RAVDESS, while the recognition rates at the diagonal elements are the highest, the recognition accuracy of positive emotions is higher than those of negative ones.

### *E. Recognition Positive and Negative Emotions*

Studies in the literature suggest that time functionality is more obvious on negative and positive emotion categories [8]. RAVDESS has a reasonable number of positive and negative emotions. As a result, we report the performance on its negative emotions (namely: sadness, anger, and fear), and positive emotions (namely: calm and happiness). Fig. 5 provides the recognition speed over time for these two categories using our framework. Interestingly, we noticed that the recognition scores increase faster for positive emotions than for the negative ones. Indeed, the figure shows that time has a more rapid impact on the negative emotions, while the recognition plateau is reached earlier for positive emotions.

In other words, on average, 2 and 3 seconds are enough for reaching the best detection accuracy for positive and negative emotions, respectively. Another interesting outcome is that positive emotions are better recognized compared with the negative ones. This is due to the fact that video-modality has better performance than audio-modality (as it can be noticed in Fig. 4), which causes this gap, since visual information is more powerful in the detection of positive emotions [19].

### *F. Conclusion*

In this paper, we proposed an end-to-end multimodal and temporal DML for AVER. The novel methodology embeds audio-visual cues diachronically, taking the advantages of time-windows of emotion display. The proposed incremental perception, based on the acquired representations from the framework, shows its efficiency at modeling the temporal context of multimodal emotion recognition. The obtained results are significantly better than the baseline results and set a new state-of-the-art performance for CREMA-D and RAVDESS. The future direction of this work includes an evaluation on a dataset with spontaneous emotion expressions, in a heterogeneous manner for multimodal inference.


### ACKNOWLEDGMENT

This work has been funded through H2020-MaTHiSiS project under Grant Agreement No. 687772. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Nvidia TITAN XP GPUs used for this research.

## REFERENCES


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

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[4] C.-H. Wu, Y.-M. Huang, and J.-P. Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.

[5] B. H. Y. B.G.Lee, T.W.Chong and B.Kim, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS*, 2017.

[6] S. Cosentino, E. I. Randria, J.-Y. Lin, T. Pellegrini, S. Sessa, and A. Takanishi, "Group emotion recognition strategies for entertainment robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 813–818.

[7] L. Y. Mano, B. S. Faiçal, L. H. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, P. Geraldo Filho, G. T. Giancristofaro, G. Pessin, B. Krishnamachari *et al.*, "Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition," *Computer Communications*, vol. 89, pp. 178–190, 2016.

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

[9] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[10] V. Aubergé and M. Cathiard, "Can we hear the prosody of smile?" *Speech Commun.*, vol. 40, no. 1-2, pp. 87–97, Apr. 2003.

[11] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," *arXiv preprint arXiv:1808.05561*, 2018.

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

[13] 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.

[14] F. Grosjean, "Gating," *Language and cognitive processes*, vol. 11, no. 6, pp. 597–604, 1996.

[15] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[16] Z. Xu, K. Q. Weinberger, and O. Chapelle, *Distance metric learning for kernel machines*, arXiv, 2012.

[17] J. Lu, J. Hu, and J. Zhou, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76–84, 2017.

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

[19] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[21] Y. Li, M. Yang, and Z. M. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[23] J. Yu, X. Yang, F. Gao, and D. Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE transactions on cybernetics*, vol. 47, no. 12, pp. 4014–4024, 2016.

[24] M. Abavisani and V. M. Patel, "Deep multimodal subspace clustering networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1601–1614, 2018.

[25] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, 2016, pp. 1857–1865.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[27] J. Lee, S. Abu-El-Haija, B. Varadarajan, and A. P. Natsev, "Collaborative deep metric learning for video understanding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. New York, NY, USA: ACM, 2018, pp. 481–490.

[28] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, 2017.

[29] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[30] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[31] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[32] R. Beard, R. Das, R. W. Ng, P. K. Gopalakrishnan, L. Eerens, P. Swietojanski, and O. Miksik, "Multi-modal sequence fusion via recursive attention for emotion recognition," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018, pp. 251–259.