MVP: Multimodal Emotion Recognition based on Video and Physiological Signals

Valeriya Strizhkova[1], Hadi Kachmar[1], Hava Chaptoukaev[2],
Raphael Kalandadze[3], Natia Kukhilava[3], Tatia Tsmindashvili[3],
Nibras Abo-Alzahab[1], Maria A. Zuluaga[2], Michal Balazia[1],
Antitza Dantcheva[1], François Brémond[1], Laura M. Ferrari[1,4]

[1] Centre Inria d'Université Côte d'Azur, Sophia-Antipolis, France
[2] EURECOM, Sophia-Antipolis, France
[3] Georgian Technical University, Tbilisi, Georgia
[4] Scuola Superiore Sant'Anna, Pontedera, Italy.

Abstract. Human emotions entail a complex set of behavioral, physiological and cognitive changes. Current state-of-the-art models fuse the behavioral and physiological components using classic machine learning, rather than recent deep learning techniques. We propose to fill this gap, designing the Multimodal for Video and Physio (MVP) architecture, streamlined to fuse video and physiological signals. Differently then others approaches, MVP exploits the benefits of attention to enable the use of long input sequences (1-2 minutes). We have studied video and physiological backbones for inputting long sequences and evaluated our method with respect to the state-of-the-art. Our results show that MVP outperforms former methods for emotion recognition based on facial videos, EDA, and ECG/PPG. The code will be available on GitHub.

1 Introduction

The recognition of human behavior from video data has been deeply investigated in the affective computing area [7]. What is limiting the performance of computer vision algorithms is the recognizability and the meaning of visual patterns. Facial expressions visible from cameras do not always correlate with the experienced emotions, which can lead to incorrect emotion recognition, e.g., a person may smile without being happy. In order to improve emotion recognition performance, *multimodal learning* has been introduced. Most of the current research on multimodal emotion recognition focuses on the combination of video, audio, and language cues. [6, 11]. Such cues represent voluntary expressions of emotions, i.e., the *behavioral* component. On the other hand, the *physiological* component is largely involuntary and is directly activated when a person is emotionally triggered [24]. When an emotional trigger appears, a change in the physiological pattern is inevitable and detectable [13]; e.g., the heart rate increases if a person sees a snake. Thanks to their involuntary nature and to the ease of recording, peripheral signals as the electrocardiography (ECG) or
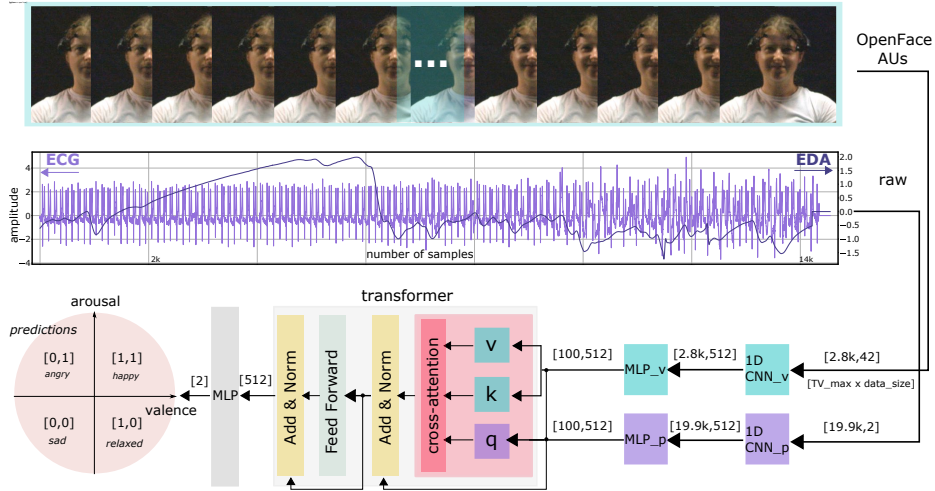
Fig. 1: Multimodal Emotion Recognition based on Video and Physiological Signals (MVP) architecture. Video features and raw physiological signals are input as full long sequences into the model to predict binary valence and arousal. The cross-attention transformer is used to fuse multiple modalities.

photoplethysmogram (PPG), for the heart activity, and the electrodermal activity (EDA), for to the skin conductivity, have been exploited for the purpose of recognizing human emotions [15, 23, 28].

While unimodal video and physiological signals have been used to recognize emotions, very few works combine the two kind of data [5, 19, 25]. A recent work from 2023, proposes to fuse traditional features, classifying them by SVM and k-NN [19]. Similarly, in a recent dataset for stress identification [5], the baselines for both unimodal and multimodal experiments are conducted based on SVM/MLP using hand-crafted features for both video and physiological signals. The state-of-the-art approaches utilize old-fashioned machine learning pipelines. This is probably due to the small number and small size (max 60 subjects) datasets available. We propose to fill this gap, designing a model that exploits the benefits of the transformer architecture *attention* [29]. The idea is to enable the use of long input sequences (1-2 minutes), in the context of *emotion recognition*, with video and physiological data, in order to extract complementary features.

First, we study how to efficiently input the video data. We compare two approaches, the DL features extracted through VideoMAE [27] and the action units (AUs), extracted through OpenFace [3]. VideoMAE and its modification [4] have just recently showed to achieve state-of-the-art results on multiple tasks, comprehending facial expression recognition. The main limitation of the available approaches is related to the input length. Typically short videos of few seconds are used. An open question is if and how these models can scale on longer videos. Here we adapt the Hugging Face implementation of VideoMAE to long

(1-2 minutes) videos, and we compare these results with the traditional approach of AUs extraction. AUs are fine-grained facial muscle movements [9], typically extracted through neural networks pre-trained on large sets of facial data. Each AU is associated with a specific set of facial muscles. As this is the first study on scaling VideoMAE for long emotion recognition videos, we assessed multiple configurations, on two datasets (AMIGOS and DEAP).

Second, we study how to efficiently input physiological data. The state-of-the-art is based on 1D-CNN and unimodal transformer for the recognition of emotions from raw physiological data [30,31]. One main limitation of these methods is related to the input signals, which are cut into 10-seconds segments. Here we improve the physiological backbone permitting the input of the full sequences, thus enabling attention to find correlation in both short and long range dependencies. Finally, we propose a multimodal transformer-based architecture, the Multimodal for Video and Physio (MVP), to efficiently couple the behavioral and physiological components for the emotion recognition task (Figure 1). The model combines facial video features with ECG/PPG and EDA. Deviating from other multimodal transformer-based architectures for video and language data [6,11,34], which use as input short subsequences (4-10 seconds), the MVP architecture is developed to handle long input sequences (1-2 minutes) of complementary behavioral and physiological data. Summarized contributions:

1. We comprehensively study the video and physiological backbones for inputting long sequences.
2. We propose MVP, a new affective computing architecture, streamlined to fuse video and physiological signals.
3. Our experimental results suggest that MVP outperforms former methods for emotion recognition based on facial videos, EDA, and ECG/PPG.

2   Related Work

2.1   Multimodal Emotion Recognition

Table 1 shows the state-of-the-art emotion recognition methods. In the case of video, audio and language, transformer-based architectures have been proposed. The data are typically input into transformer as features [31,33,34]: Action Units for the video (e.g., from OpenFace), audio features (e.g., Mel spectrogram from Librosa or Covarep), and language features (e.g., from GloVe or BERT). Each modality is used as input into self-attention layers, which allow the model to attend to different parts of the input sequence and capture long-range dependencies. Then the multimodal part takes place, performing cross-attention across modalities. The cross-attention can be done in multiple ways. The cross-modal fusion has been developed on pairwise representations [12]. Here two couples of the modalities (i.e., language + audio and language + video, with language as the primary modality) enter a complementary module where bimodal fusion is done at the Multi-Head Attention level. Here, each modality has its own representation to preserve modality-specific features. A joint-encoding has been also used

Table 1: State-of-the-art multimodal emotion recognition methods. V means visual, A means audio, L means language, P means physiological. We describe visual and physiological features in detail, but do not describe audio and linguistic features since our pipeline is based on video and physiological modalities. M-SENA is a platform with the state-of-the-art ER methods, from which we reported here the main characteristics of top three methods.

Year	Framework	Modalities	Visual (V) and Physio (P) Features
2019	UDLBV [25]	V, P	VGG-16 (V)
2020	TBJE [34]	V, A, L	R(2+1)D-152 outputs (V)
2021	BBMF [11]	V, A, L	Action Units (V)
2022	M-SENA [17]	V, A, L	Action Units (V) Facial Landmarks (V) Eye Gaze (V)
2022	MultiTransformer [30]	P (ECG, EEG)	1D-CNN + Transformer (P)
2023	EHCI [19]	V, P	Hand-Crafted (V, P)
2023	StressID [5]	V, P (ECG, EDA)	Hand-Crafted (V, P)

at the level of the Multi-Head Attention, to encode information from all the modalities, producing a unified representation [6]. The best result is achieved with a bimodal model using the linguistic + acoustic cues. Notably, in both cases [6, 12], the language cue is the most representative. This can be due to the fact that annotators mainly worked on the transcript data. One of the major limitation of these works is related to the unique use of the behavioral component, i.e., video, audio and language data.

Few attempts have been made to build a multimodal framework that can exploit both the behavioral and physiological components. These works rely on late fusion of classical features. In the UDLBV [25] a CNN architecture (VGG-16 pretrained on ImageNet) and hand-crafted (HC) features are concatenated in a first step, after which PCA is applied, for then passing the features to an LSTM (only for the DEAP dataset) and finally classify with extreme learning machines (ELM). In EHCI [19] traditional features as spectral, HC, and local binary patterns for facial images are late fused and classified through SVM and k-NN.

Our aim is to propose a transformer-based architecture that fuses video and physiological data through cross-attention. The idea is to improve the emotions recognition by combining these two complementary signals, which represent the behavioral and the physiological components of emotions.

Video cues. State-of-the-art video-based emotion recognition methods use two main types of facial video features: (i) classical features like action units (AUs) and facial landmarks [11, 17] and (ii) general features extracted through DL techniques [33, 34]. We compare these two approaches: (i) the use of AUs, ex-

tracted through OpenFace [3], and (ii) general features extracted through Video-MAE [27]. AUs are fine-grained facial muscle movements [9], typically extracted through neural networks pre-trained on large sets of facial data. These methods have been largely used showing stable and reliable results. VideoMAE and its modifications [4, 32] are new pre-training methods, providing the best results on many tasks, including video-based facial expression recognition. Masked auto-encoders (MAE) are self-supervised learners (SSL) trained to reconstruct the missing masked random patches of the input video. The use of VideoMAE in the emotion recognition field is very recent, therefore we report our adaptation to the challenging set-up of emotion recognition with long videos.

Physiological cues. Peripheral physiological data are largely involuntary [24], mediated by the autonomus nervous system, which is directly activated when a person is emotionally triggered. This means that when an emotional activation is present, a change in physiological patterns is inevitable and detectable [13]. Peripheral physiological data, as ECG/PPG and EDA, have been exploited to derive reliable information for emotion recognition [8]. The use of machine learning on physiological data is quite recent, up to 5-6 years ago computational models were the most used. In the last years, classical machine learning pipelines, with models as Multilayer Perceptron (MLP), Support Vector Machine (SVM) and random forest trained on HC features [21] became more popular. The first approach exploiting deep learning architectures has seen the translation of physiological time-series data into 2D images, representing spectrogram or power spectral density, which are then input into CNN architectures (as ResNet or VGG) pretrained on different datasets (as ImageNet) [2, 7, 10]. GNNs have also been used [14] and two stacked convolutional autoencoders has been proposed to combine ECG and EDA [21]. Now, the state-of-the-art method in the field is based on 1D-CNN and self-attention transformers for the recognition of emotions from raw physiological data [30, 31]. The results have been proved only for the AMIGOS dataset. In ERMS [30] the two signals are used as time-series and the prediction is done after the concatenation of the last hidden layers of unimodal transformers. One main limitation of all these methods is related to the input signals, which are cut segments of 10-seconds. Then, or random 10s are taken for each sample, or each sample is used and the output of the features extractor is averaged. The advantage of using 10-seconds subsequences is that the network can work with a small and fixed input, resulting in a higher efficiency. This approach does not consider two factors. The ECG and EDA vary on different time scales, from few ms to tens of seconds/minute. The annotations are weak; a subject may express the labeled emotion for some seconds, few times in the whole recording, and being neutral the rest of the time. We advance the state-of-the-art permitting the input of the full sequences, thus enabling attention to find correlation in both short and long range dependencies.

We compare our approach with with state-of-the-art methods for the input of physiological data [31] and for the combination of multimodal signals [30].

Table 2: Video datasets used in this study. KAD stays for Kinetics-400 & AMIGOS & DEAP.

Dataset	# Video clips	Average clip duration	Source
Kinetics-400	306245	10s	YouTube
AMIGOS	640	80s	Lab
DEAP	880	70s	Lab
KAD	307765	-	Multi-source

3 Method

3.1 VideoMAE for Long Input Videos

We use the Hugging Face implementation of VideoMAE, with the ViT-B encoder of 12 layers, hidden size of 768, and 87M parameters. During pre-training, the input video is masked with a ratio of 90% and fed into the encoder, which outputs the latent representations. The shallow decoder then takes the latent representations from the encoder and reconstructs the input videos. During fine-tuning, the pre-trained encoder is learned to predict binarized valence and arousal from original (i.e. not masked) videos. We perform experiments on two datasets, AMIGOS [18] and DEAP [16]. We implemented and tested multiple configurations of pre-training strategy. (i) Self-supervised pre-training on Kinetics-400 and fine-tuning on the target dataset (i.e. AMIGOS or DEAP). (ii) Self-supervised pre-training on Kinetics-400, plus AMIGOS and DEAP, and fine-tuning on the target dataset (see Figure 2). (iii) Self-supervised pre-training on Kinetics-400, plus an intermediate step of supervised pre-training on Kinetics-400 and one of the two target datasets and fine-tuning on the other target dataset. Table 2 shows input video datasets used. Figure 2 represents the pipeline of VideoMAE pre-training, with masked input video and reconstruction task, and the fine-tuning step to predict valence and arousal.

3.2 Selected Backbones for Features Extraction

Vision backbone. Our video backbone is composed of 1D-CNN$_v$ followed by MLP$_v$. It takes as input the AUs matrix of the full input video and outputs embeddings that are used as tokens in transformer.

Physiological backbone. Our physiological backbone extracts temporal features from the entire raw physiological data using 1D-CNN$_p$ and MLP$_p$ layer. We expanded the use of the sole 1D-CNN$_p$ [31], adding the MLP$_p$, to enable the handling of long sequences. The extracted features are used as tokens in transformer.
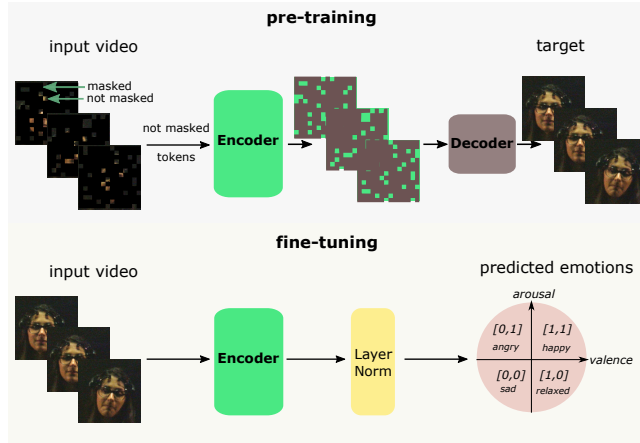
Fig. 2: Pre-training and fine-tuning steps of VideoMAE in emotion recognition. In the pre-training step the autoencoder reconstructs the masked input video. During the fine-tuning, the pre-trained encoder is fine-tuned to predict binarized valence and arousal from the original not masked videos.

3.3   MVP Architecture

Figure 1 shows the MVP architecture. Here, one input trial is exemplified, consisting of a video sequence SV_{ij} and a physiological sequence SP_{ij}, of shapes $TV_{max} \times 42$ and $TP_{max} \times 2$, respectively. In our experiments, we use a batch size of 8, so at each training iteration we randomly choose 8 trials and perform all the calculations for this batch. The input data enters the video and physiological backbones. For the video input, the zero-padded matrix with AUs enters 1D-CNN$_v$ followed by MLP$_v$. For the physiological data, the zero-padded matrix with concatenated ECG and EDA enters 1D-CNN$_p$ followed by MLP$_p$. 1D-CNN$_v$ and 1D-CNN$_p$ extract features from input data without changing the time dimension. MLP$_v$ and MLP$_p$ reduce the time dimension without changing the feature dimension. 1D-CNN$_p$ takes as input a raw physiological sequence of dimension [19.9k, 2] and outputs a feature sequence of dimension [19.9k, 512]. Then MLP$_p$ takes as input the output from 1D-CNN$_p$ and outputs a compact representation [100, 512].

After that, the outputs from 1D-CNNs are used as input into transformer with cross-attention for fusing video with physiological signals. The MVP method uses mid-fusion fusion, where video modality is input as keys (k) and values (v) and physiological one as queries (q) (Equation 1).

$$\text{Attention}(Q_p, K_v, V_v) = \text{Softmax}\left(\frac{Q_p K_v^T}{\sqrt{d_k}}\right) \cdot V_v \qquad (1)$$

where Q_p is the query from the physiological data and K_v and V_v are the keys and the values from the video data. transformer has 8 attention heads and 8
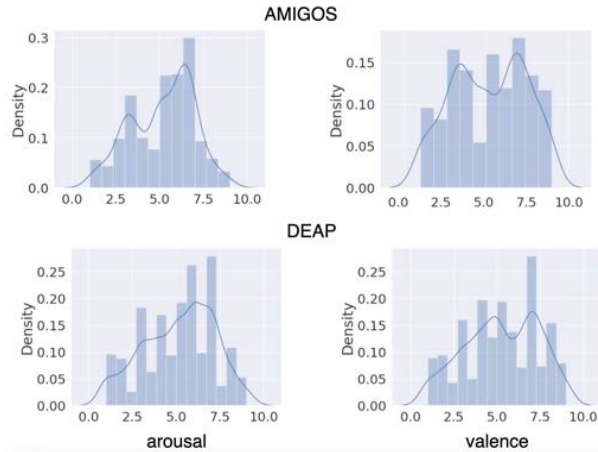
Fig. 3: Label distribution of AMIGOS and DEAP datasets.

attention layers and it predicts two outputs: valence and arousal, binarized, as reported in Figure 1. We use binary cross entropy (BCE) to calculate the loss. During training the loss is backpropagated through the neural network layers to calculate the gradient for each weight.

3.4 Data Handling

The entire sequence of video and physiological data is used, with a length that can vary between 60s and 155s. As the adopted datasets contain multiple trials from different subjects, the input data for a trail i of a subject j is composed of two matrices: SV_{ij} for video and SP_{ij} for physiological signals, with dimensions $[TV_{ij}, 42]$ and $[TP_{ij}, 2]$, respectively, where TV_{ij} and TP_{ij} correspond to lengths of the sequences and 42 and 2 are input data dimensions of each modality.Video and physiological signals are sampled at different rates, therefore TV_{ij} and TP_{ij} are different (e.g., for the AMIGOS dataset TV_{max} is 2.8k and TP_{max} is 19.9k, corresponding to 155s sampled at 18fps and 128Hz, for the video and the physiological data, respectively). We extract 42 visual features from each video frame and use the sequence of AUs and eye gaze as video representation. AUs are fine-grained facial muscle movements [9], each of which relates to a subset of extracted facial landmarks [20]. From each frame we extract the following AUs: $1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28,$ and 45, by means of the Open-Face library [3]. Each AU is described in two ways: presence, if AU is visible in the face, and intensity, how intense is the AU on a 5-point scale. The eye gaze corresponds to two gaze direction vectors of each eye. The physiological data has a dimension of 2 at each timestamp, corresponding to the $2D$ time series, made of ECG and EDA raw signals.

Before using the data as input to our model, we perform the following pre-processing. We find the longest sequence and use its TV_{max} and TP_{max} lengths

to construct input matrices for all the trials in the dataset. Each pair of input sequences SV_{ij}, SP_{ij} is zero-padded so that all input matrices with AUs and physiological signals have sizes $TV_{max} \times 42$ and $TP_{max} \times 2$, respectively.

4   Experiments

4.1   Emotion Recognition Labels

Emotion classification is typically approached by measuring arousal and valence based on the emotion circumplex model [22]. Valence refers to the degree of positive or negative affect associated with an emotion. Arousal refers to the degree of activation, engagement associated with an emotion. The valence and arousal dimensions have been used to categorize emotions into quadrants (see Figure 1): high valence/high arousal (e.g., excitement, happiness), high valence/low arousal (e.g., relaxation), low valence/high arousal (e.g., anger), and low valence/low arousal (e.g., boredom, tiredness). In the classification task these labels are typically extracted by the Self-Assessment Manikin (SAM) scale reported by each subject. SAM is a pictorial tool for rating valence and arousal on a scale of 1 to 9. A subject selects the pictogram that best corresponds to its emotional state at the end of each elicitation video. Therefore, such labels are more subjective and referred to the whole session, that typically last between 1 and 2 minutes. In Figure 3, the label distribution of the used datasets is reported.

4.2   Datasets and Pre-processing

The AMIGOS [18] and the DEAP [16] datasets are available online and comprise both video and physiological data. The recordings are performed in a laboratory setting and the participants are instructed to watch emotional videos and then respond to the Self-Assessment Manikin scale. The AMIGOS dataset includes the recording of 40 subjects, who rated 16 movie video clips for valence, arousal, and other measures (i.e., dominance, liking, familiarity and basic emotions). The DEAP dataset includes the recording of 32 subjects, who rated 40 music video clips for valence, arousal and other measures. We use the facial videos and EDA + ECG for AMIGOS, and the facial videos and EDA + PPG for DEAP, preprocessed as follow.

For VideoMAE, the input videos are cut into smaller clips of 4 seconds, from which 16 frames are taken and cropped using facial landmarks generated by OpenFace [26]. Two different cropping strategies were evaluated (Figure 4). The larger one captures the entire face and a bit of background, ensuring the face remains visible even when the subject head is in motion, the smaller one captures exclusively the face, causing occasionally and partially cropping out. We experimented the two cropping strategies, achieving higher performances (7-9%) with the larger crops; thereafter this is the set-up adopted in this work.

For the AUs, we use OpenFace [1] to extract AUs, as it is a stable and largely adopted library in the community. 42 AUs and eye gaze direction features are extracted for each frame.
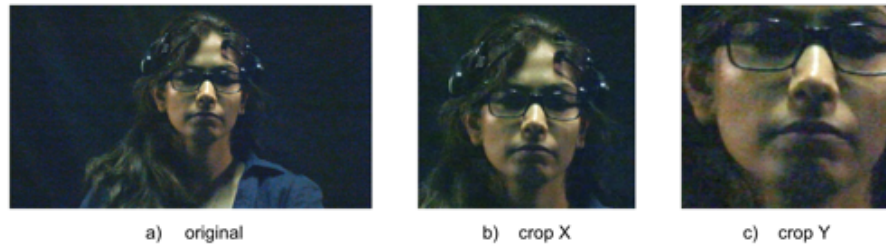
a)   original   b)   crop X   c)   crop Y

Fig. 4: Crop comparison on the Amigos dataset. The larger crop X captures the entire face and a bit of background, the smaller crop Y captures exclusively the face. Larger crops give higher F1-score.

For the physiological data, we downsampled to 128Hz, then we filtered the signals with Butterworth filter plus powerline filter (to remove the 50Hz). The Butterworth filters used are: 3th order band-pass with [0.5,8]Hz band for PPG, 5th order high-pass with a 0.5Hz cutoff for ECG, 4th order low-pass with a 3Hz cutoff for EDA. In the AMIGOS dataset, we cut the first 1s in all the physiological signals, as artefacts are present in most of the recordings (probably due to the device connection). Since the signal sequences have different lengths, we pad the shorter sequences with zeros, so that all input sequences have the same length. The raw data is normalized with zero-mean and unity-variance before entering the pipeline.

4.3   Experimental Set-up

We use 5-fold cross validation to split data into train and test subject independently. All trials of one subject are either in the train or in the test split. We use the weighted F1-scores as metric, as the datasets are imbalanced, reporting the mean and standard deviation.

We split valence and arousal continuous labels using the thresholds of 4.5 and 5 for AMIGOS and DEAP, respectively. If the continuous label is less or equal than the threshold, the binarized label is 0 and 1 otherwise. In Figure 3 the labels distribution of both AMIGOS and DEAP datasets are reported.

5   Results and Discussions

5.1   Input Features

Our model fuses video with physiological signals. To extract features from each modality, we use the models that best fit the input data. Here we explain our motivation in selecting feature extraction backbones for facial videos and physiological signals (ECG and EDA).

Table 3: Valence prediction from raw videos using self-supervised VideoMAE pre-training and supervised pre-training. KAD is Kinetics-400 & AMIGOS & DEAP.

SSL pre-training data	Supervised pre-training data	F1-score weighted
Kinetics-400	AMIGOS	49 ± 5
KAD	AMIGOS	50 ± 4
Kinetics-400	Kinetics-400 & AMIGOS	51 ± 6
Kinetics-400	KAD	$\mathbf{59 \pm 5}$

(a) 5-fold cross validated F1-score for the AMIGOS dataset.

SSL pre-training data	Supervised pre-training data	F1-score weighted
Kinetics-400	DEAP	52 ± 4
KAD	DEAP	55 ± 7
Kinetics-400	Kinetics-400 & DEAP	53 ± 5
Kinetics-400	KAD	$\mathbf{60 \pm 6}$

(b) 5-fold cross validated F1-score for the DEAP dataset.

VideoMAE for emotion recognition. The results in Tables 3a, 3b are related to the valence prediction for the AMIGOS and DEAP datasets using the VideoMAE self-supervised pre-training. The results show that the most effective approach is the three steps strategy, where, after the SSL pre-training on Kinetics-400, a supervised pre-training step is added before fine-tuning on the target dataset.

While increasing the amount of data in the self-supervised pre-training phase does not achieve the highest weighted F1-score, it is still more effective than solely relying on Kinetics-400 for supervised pre-training or not using it at all. Additionally, when the supervised pre-training step is added, the model adapts faster. For instance, with DEAP, the model requires only 3 epochs to fine-tune, as opposed to the initial 10. This further supports the intuition that leveraging datasets from the same domain can significantly benefit the learning process.

Selection of the video backbone. We compare two approaches: (i) the use of AUs extracted through OpenFace, and (ii) DL features extracted through VideoMAE [27]. We compare the results on the AMIGOS dataset in Table 4a. The two approaches give similar results, with AUs outperforming VideoMAE features. We selected AUs features as they give higher F1-score, are more reliable, explainable and requires less memory for training. A possible explanation for their performance is that the model used to extract AUs is pre-trained on a large set of face data, so the predictions incorporate extra knowledge, providing stable results. In this context, stable means that the extracted features are meaningful under different lighting conditions, resolutions, and head angles. Indeed, on the

Table 4: Comparison of state-of-the-art emotion recognition methods with the proposed MVP for unimodal and multimodal emotion recognition. ECG+EDA means concatenated raw ECG and EDA used as one modality. Video input is raw videos for VideoMAE and AUs for other methods. HC means hand-crafted.

Method	F1-score weighted					
	ECG+EDA		video		ECG+EDA+video	
	arousal	valence	arousal	valence	arousal	valence
SVM (HC feat) [5]	51 ± 5	60 ± 6	44 ± 8	59 ± 3	49 ± 8	64 ± 3
MLP (HC feat) [5]	57 ± 4	63 ± 1	53 ± 5	64 ± 5	55 ± 5	63 ± 3
UniTransformer (adapted) [31]	54 ± 4	59 ± 4	—	—	—	—
MultiTransformer (adapted) [30]	—	—	—	—	56 ± 6	61 ± 3
VideoMAE [27]	—	—	53 ± 6	59 ± 5	—	—
UniAUTransformer (ours)	—	—	56 ± 3	63 ± 4	—	—
UniPhysioTransformer (ours)	56 ± 3	62 ± 4	—	—	—	—
MVP (10s cut input, ours)	—	—	—	—	55 ± 5	60 ± 6
MVP (ours)	—	—	—	—	**58 ± 6**	**66 ± 4**

(a) 5-fold cross validated F1-score for the AMIGOS dataset.

Method	F1-score weighted					
	ECG+EDA		video		ECG+EDA+video	
	arousal	valence	arousal	valence	arousal	valence
MLP (HC feat) [5]	52 ± 3	53 ± 2	51 ± 2	54 ± 3	52 ± 2	55 ± 6
VideoMAE [27]	—	—	52 ± 3	60 ± 6	—	—
MVP (ours)	53 ± 6	54 ± 4	54 ± 4	57 ± 7	**55 ± 5**	**61 ± 4**

(b) 5-fold cross validated F1-score for the DEAP dataset.

AMIGOS dataset, where the videos are quite dark and low resolution, AUs improve the results even more.

Selection of the physiological backbone. We compare the state-of-the-art unimodal transformer for physiological data [31] and our physiological backbone. The unimodal transformer backbone uses 10s subsequences and 1D-CNN while our backbone is made of 1D-CNN + MLP and data is input as whole sequences. Table 4a shows that our proposed approach outperforms the state-of-the-art by around 3 percentage points in predicting valence.

5.2   Multimodal Emotion Recognition

The need of combining video and physiological data through a novel DL architecture is investigated. We compare the MVP method with classical machine learning approaches and more recent transformer-based architectures proposed

for both unimodal and multimodal data. The classical machine learning pipeline used here exploits hand-crafted (HC) features as input to Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM) to classify valence and arousal [5]. All transformer models are used with adaptations to ensure fair comparison. In UniTransformer we combine ECG and EDA, whereas the original UniTransformer uses only ECG as input. In MultiTransformer, we add AUs as video component, using ECG+EDA as one input and AUs as the second one. The original MultiTransformer uses ECG and EEG. The main difference between MultiTransformer and MVP is the fusion. In MultiTransformer late fusion is used, while MVP uses mid-fusion. For VideoMAE we have studied and adapted the method to the case of long facial videos. Tables 4a, 4a show that transformers outperform classical machine learning. And the use of mid-fusion and a temporal model able to input the full sequence, as is done in MVP, gives the best results. This is due to the capability of cross-attention to find long-range dependencies between different kinds of input data.

5.3   Success Case Analysis

We conduct case studies to show videos and physiological sequences in which the proposed MVP method outperforms state-of-the-art competitors. We compare our MVP with MLP proposed by StressID for multimodal fusion [5]. Table 4b shows that MVP outperforms MLP on the DEAP dataset. Figures 5a and 5b demonstrate the sequences where MVP gives better results than the competitor. We show both physiological signals and video frames of our success cases. Figure 5a shows the case where the true arousal is 1, our MVP method predicts 1, and the competitor predicts 0. The video frames above show that the participant is initially neutral, and after a few seconds they smile, while at the same time the physiological signals change. The result demonstrates that MVP better captures the dynamics of the sequence. Figure 5b shows the case where the true valence is 0, our MVP method predicts 0, and the competitor predicts 1. The video frames above show that the participant remains neutral throughout the experiment, without any changes except blinking while the physiological signal changes. This success case shows that MVP better fuses multiple modalities than MLP.

6   Conclusions and Future Work

We propose a new method for combining complementary data, the behavioral and physiological component of emotions. The method relies on the relevance of inputting long full sequences, exploiting attention. Regarding the video input, AUs are performing better than deep learning features in extracting facial representations for emotion recognition. Nevertheless, a strategy that includes similar data distribution in the pre-training step shows improved results for deep learning features. Our experimental results showcase that fusing video and physiological signals outperforms each modality individually. MVP achieves state-of-the-art results in the challenging field of multimodal emotion recognition with
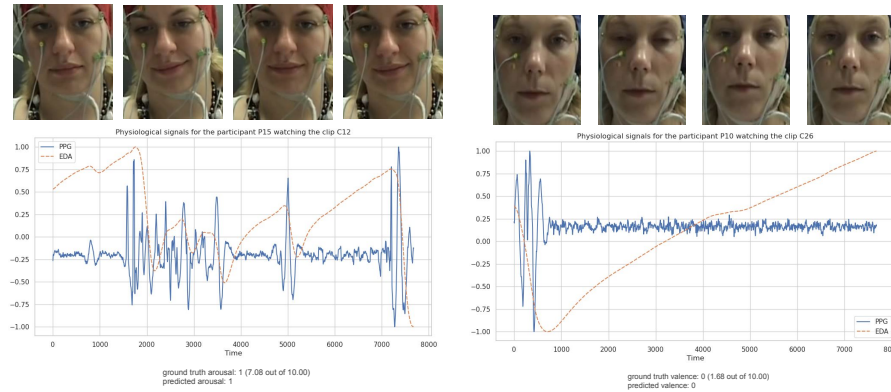
(a) The success case for the arousal prediction.   **(b)** The success case for the valence prediction.

Fig. 5: The success cases for the arousal and valence prediction on the DEAP dataset.

video and physiological data, where small datasets are available. With this work we aim to foster research in the field, to improve the understanding of human emotions expressed by behavioral and physiological signals.

Future works involve exploiting ECG and EDA data as two separated modalities, fused by a transformer model. We will develop a dedicated pre-training technique, in order to increase recognition capabilities of the model and to bring insight into physiological and behavioral modalities.

References

1. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science (2016) 9

2. Bagherzadeh, S., Maghooli, K., Shalbaf, A., Maghsoudi, A.: Recognition of emotional states using frequency effective connectivity maps through transfer learning approach from electroencephalogram signals. Biomedical Signal Processing and Control **75**, 103544 (2022) 5

3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018) 2, 5, 8

4. Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofighi, H., Haffari, R., Hayat, M.: Marlin: Masked autoencoder for facial video representation learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 5

5. Chaptoukaev, H., Strizhkova, V., Panariello, M., Dalpaos, B., Reka, A., Manera, V., Thummler, S., Ismailova, E., W., N., Bremond, F., Todisco, M., Zuluaga, M.A., Ferrari, L.M.: StressID: a multimodal dataset for stress identification. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=qWsQi9DGJb 2, 4, 12, 13

6. Delbrouck, J.B., Tits, N., Brousmiche, M., Dupont, S.: A transformer-based joint-encoding for emotion recognition and sentiment analysis. Annual Meeting of the Association for Computational Linguistics (ACL) (2020) 1, 3, 4

7. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 467–474 (2015) 1, 5

8. Egger, M., Ley, M., Hanke, S.: Emotion recognition from physiological signal analysis: A review. Electronic Notes in Theoretical Computer Science **343**, 35–55 (2019) 5

9. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978) 3, 5, 8

10. Elalamy, R., Fanourakis, M., Chanel, G.: Multi-modal emotion recognition using recurrence plots and transfer learning on physiological signals. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–7 (2021). https://doi.org/10.1109/ACII52823.2021.9597442 5

11. Han, W., Chen, H., Gelbukh, A., Zadeh, A., philippe Morency, L., Poria, S.: Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. ACM International Conference on Multimodal Interaction (ICMI) (2021) 1, 3, 4

12. Han, W., Chen, H., Gelbukh, A., Zadeh, A., philippe Morency, L., Poria, S.: Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. ACM International Conference on Multimodal Interaction (ICMI) (2021) 3, 4

13. Jerritta, S., Murugappan, M., Nagarajan, R., Wan, K.: Physiological signals based human emotion recognition: a review. In: 2011 IEEE 7th international colloquium on signal processing and its applications. pp. 410–415. IEEE (2011) 1, 5

14. Jia, Z., Lin, Y., Wang, J., Feng, Z., Xie, X., Chen, C.: Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1047–1056 (2021) 5

15. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H., Koo, B.H.: Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry investigation **15**(3), 235 (2018) 2

16. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing **3**(1), 18–31 (2011) 6, 9

17. Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., Gao, K.: M-sena: An integrated platform for multimodal sentiment analysis. Annual Meeting of the Association for Computational Linguistics System Demonstration Track (ACL) (2022) 4

18. Miranda-Correa, J.A., Abadi, M.K., Sebe, N., Patras, I.: Amigos: A dataset for affect, personality and mood research on individuals and groups. IEEE Transactions on Affective Computing (2021) 6, 9

19. Moin, A., Aadil, F., Ali, Z., Kang, D.: Emotion recognition framework using multiple modalities for an effective human–computer interaction. The Journal of Supercomputing **79**(8), 9320–9349 (2023) 2, 4

20. Perveen, N., Mohan, C.K.: Configural representation of facial action units for spontaneous facial expression recognition in the wild. In: VISIGRAPP (4: VISAPP). pp. 93–102 (2020) 8

21. Ross, K., Hungler, P., Etemad, A.: Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data. Journal of Ambient Intelligence and Humanized Computing (2021) 5

22. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980) 9
23. Sánchez-Reolid, R., López, M.T., Fernández-Caballero, A.: Machine learning for stress detection from electrodermal activity: A scoping review. Preprints (2020) 2
24. 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 **18**(7), 2074 (2018) 1, 5
25. Siddharth, S., Jung, T.P., Sejnowski, T.J.: Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. IEEE Transactions on Affective Computing **13**(1), 96–107 (2019) 2, 4
26. Tadas Baltrušaitis, Peter Robinson, L.P.M.: Openface: An open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (2016) 9
27. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems (NeurIPS) (2022) 2, 5, 11, 12
28. Topic, A., Russo, M., Stella, M., Saric, M.: Emotion recognition using a reduced set of eeg channels based on holographic feature maps. Sensors **22**(9), 3248 (2022) 2
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) (2017) 2
30. Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.L.: Emotion recognition with pre-trained transformers using multimodal signals. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–8. IEEE (2022) 3, 4, 5, 12
31. Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.L.: Transformer-based self-supervised learning for emotion recognition. International Conference on Pattern Recognition (ICPR) (2022) 3, 5, 6, 12
32. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 5
33. Zhang, T., Liu, C., Liu, X., Liu, Y., Meng, L., Sun, L., Jiang, W., Zhang, F., Zhao, J., Jin, Q.: Multi-task learning framework for emotion recognition in-the-wild. European Conference on Computer Vision Workshop (ECCVW) (2022) 3, 4
34. Zhang, W., Qiu, F., Wang, S., Zeng, H., Zhang, Z., An, R., Ma, B., Ding, Y.: Transformer-based multimodal information fusion for facial expression analysis. IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2022) 3, 4