AMIGOS: A dataset for Mood, personality and affect research on Individuals and GrOupS

Juan Abdon Miranda-Correa, Student Member, IEEE, Mojtaba Khomami Abadi, Student Member, IEEE, Nicu Sebe, Senior Member, IEEE, and Ioannis Patras, Senior Member, IEEE

Abstract—We present a database for research on affect, personality traits and mood by means of neuro-physiological signals. Different to other databases, we elicited affect using both short and long videos in two settings, one with individual viewers and one with groups of viewers. The database allows the multimodal study of the affective responses of individuals in relation to their personality and mood, and the analysis of how these responses are affected by (i) the individual/group setting, and (ii) the duration of the videos (short vs long). The data is collected in two experiments. In the first one, 40 participants watched 16 short emotional videos while they were alone. In the second one, the same participants watched 4 long videos, some of them alone and the rest in groups. Participants’ signals, namely, Electroencephalogram (EEG), Electrocardiogram (EGC), and Galvanic Skin Response (GSR), were recorded using wearable sensors. Frontal, full-body and depth videos were also recorded. Participants have been profiled for personality using the Big-five personality traits, and for mood with the baseline Positive Affect and Negative Affect Schedules. Participants emotions have been annotated with both, self-assessment of affective levels (valence, arousal, control, familiarity, like/dislike, and selection of basic emotion) felt by the participants during the first experiment, and external-assessment of participants’ levels of valence and arousal for both experiments. We present a detailed correlation analysis between the different scores of personality, mood and affect. We also present baseline methods and results for single-trial classification of valence and arousal, and for single-trial classification of personality traits, mood and social context (alone vs group), using EEG, GSR and ECG and fusion of modalities for both experiments. The database has been made publicly available.

Index Terms—Emotion Classification, EEG, Physiological signals, Signal processing, Personality traits, Mood, Affect Schedules, Pattern classification, Affective Computing.

1 INTRODUCTION

Affective computing aims at detecting human emotional cues and synthesize emotional responses for Human-Computer Interaction (HCI) [1]. In this field, there is an increasing interest in considering the emotional response of users when making computational decisions that could improve or modify the effects of the interaction. For example, Chanel et al [2] modified the difficulty of a video game according to the player’s predicted boredom or anxiety in order to maintain high engagement. In other example, O’Neill [3] has proposed to use models of affect in order to generate stories more likely to produce affective responses on readers. In a more dynamic scenario, movies could be emotionally adaptive, having the possibility of changing their time-line of events and conclusion according to factors such as the viewer’s emotions, personality and mood.

For these scenarios, it is very important to reliably predict such factors. Significant progress has been done on predicting the emotional (affective) state of people in response to different stimuli, such as music videos [1], short emotional videos [4], [5], and even diverse emotion elicitation methods [6], using information from different modalities (e.g. EEG, facial expression). This progress has been boosted by the availability of multimodal annotated emotional databases, which act as a benchmark for different researchers to develop their methodologies, test their theories and compare their results.

Currently, available multimodal affective databases have focused on the study of affective responses of participants in individual [1], [7], or pairs of people/limited agent configuration [8]. However, affective experiences in real life are often performed in social contexts (e.g. audiences for movies and games are normally conformed by groups of people), where individual experiences do not depend only on the user and the content, but also on the implicit and explicit interactions that can occur between the personalities, reactions, moods and emotions of other audience members. For instance Dhall et al [9] analyzed affect of groups of people in images showing that individual facial information in combination with scene information help in inferring affect conveyed by a group, but still their analysis is done in static images. Additionally, different aspects of affect and personality could be inhibited or amplified depending on whether a person is alone or accompanied. Therefore, current databases have ignored an important dimension for the study of affect.

Databases for personality research have considered information related to linguistics on written text [10], social networks activity [11], and behavior in group activities [12]. However they have largely ignored the study of both, affect and personality, through the use of physiological signals,

The first two authors contributed equally to this work. Juan Abdon Miranda-Correa and Ioannis Patras are with the School of Computer Science and Electronic Engineering, Queen Mary University of London, UK. E-mail: {j.a.mirandacorrea,i.patras}@qmul.ac.uk.
Mojtaba Khomami Abadi and Nicu Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Italy. E-mail: {khomamia,sebe}@disi.unitn.it.
which have shown to have valuable information for personality recognition [13], [14].

Therefore, there is a need of multimodal databases for the study of people’s emotions, during affective experiences when they are both alone and in a group, considering people’s affective factors such as personality and mood. The multimodal framework would benefit from the inclusion of neurological and peripheral physiological signals.

Our contribution to the field is a database of Neurophysiological Signals for Affect, Personality and Mood Research for Individuals and Groups (AMIGOS). The database consists of multimodal recordings of participants and their responses to emotional excerpts from movies. The main distinguishing characteristics of the database are: (i) participants engage with two sets of stimuli, one of short videos and one of long videos, (ii) in the first set of videos all participants engaged in the experiments alone, whereas in the second set of videos some participants engaged in the experiments alone and some of them in groups. Additionally, participants have been profiled according to their personality through the Big-Five personality traits model, and according to their mood through the Positive Affect and Negative Affect Schedules (PANAS). Participants’ recordings for the first experiment have been annotated through affective self-assessment (internal annotation) performed by each participant at the beginning of the experiment and immediately after each video. The recordings of both sets of videos (short and long) have been off-line annotated on the scales of valence and arousal, by 3 annotators (external annotation), using a method that allows the direct comparison of the affective responses from both experiments. The modalities recorded are Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR), frontal HD video, and both, RGB and depth full body videos. The recordings have been precisely synchronized to allow the study of affective responses, personality and mood from the different modalities simultaneously. The physiological signals have been recorded using commercial wearable sensors that allow more freedom for the participants than conventional laboratory equipment (e.g. Biosemi ActiveTwo) used in [1], [4], [7], and of better quality than the equipment used in [14]. The database is available to the academic community.

In this work, we first present a comparison between the internal and external annotation of valence and arousal for the short videos experiment, which indicates that external annotation is a good predictor of the affective state of participants. (ii) We show, by correlation analysis, that in the eyes of the annotators, participants seem to have low arousal in low valence moments and high arousal for high valence moments. (iii) We found significant differences in the distribution of valence and arousal between participants that were alone with respect to participants that were in groups in the long videos experiment, whereas in the short videos experiment, where every participant was alone, both sets of participants present similar distribution of valence and arousal. (iv) We found significant negative correlations between the scores of negative affect (NA) and the ones of extraversion, agreeableness, emotional stability and openness, and significant positive correlations between the scores of agreeableness and both extraversion and positive affect (PA), between consciousness and emotional stability, and between the scores of PA and the ones of arousal. Finally, (v) our method for personality traits, mood, and social context prediction based on neuro-physiological signals of short and long videos outperforms an early baseline study [13] in prediction of all dimensions.

In section 2 we first make a survey of the main multimodal databases available for affect and personality research and compare them with our database. Additionally, we present works related to the modeling and assessment of affect, personality, and mood. Section 3 presents the experimental scenarios, stimuli selection, modalities used for recording of implicit responses, and the equipment used to record the different modalities. We also provide an overview of the experimental setup for both experiments, and the methods we employed for assessment of affect, personality traits, and mood (PANAS). In Section 4 we analyze the data obtained from the different assessments. Section 5 presents our method for single trial valence and arousal recognition and our approach for personality traits, PANAS, and social context recognition parting from neuro-physiological signals, and presents and discusses the results. Finally, we conclude in section 6.

2 RELATED WORKS

2.1 Affect, Personality and Mood

Plutchik [15] has defined emotion as a complex chain of loosely connected events that begins with a stimulus and includes feelings, psychological changes, impulses to action and specific, goal-directed behavior. The most common approaches to model affect are categorical and dimensional. The first approach claims that there exists a small number of emotions that are basic and recognized universally; The most common of these models is the Six Basic Emotions model, presented by Ekman et al [16], that categorizes emotions into fear, anger, disgust, sadness, happiness and surprise. The dimensional approach considers that affective states are not independent but rather they are inter-related in a systematic way (e.g. the Plutchik’s emotion wheel [15]). Russell [17] introduced the Circumplex Model of Affect, where affective states are represented in a two dimensional space with arousal and valence as the main dimensions of emotion. Arousal refers to the degree an emotion feels active
or inactive, and valence refers to the degree an emotion feels pleasant or unpleasant.

It has been shown that affective experiences are not only defined by external situations, but they are also modulated by people’s internal factors, such as mood and personality [18]. Personality refers to stable individual characteristics, measurable in quantitative terms that explain and predict behavior [19]. A widely used model for the study of personality is the Big-Five factor model [20], which describes personality in terms of five dimensions namely Extraversion (sociable vs reserved), Agreeableness (compassionate vs dispassionate and suspicious), Conscientiousness (dutiful vs easy-going), Emotional stability (nervous vs confident), and Openness to experience (curious vs cautious). According to John [21], Extraversion implies an energetic approach toward the social and material world; Agreeableness contrasts a prosocial and communal orientation towards others with antagonism; Conscientiousness describes socially prescribed impulse control that facilitates task and goal-directed behaviour, such as thinking before acting and delaying gratification; Neuroticism contrasts emotional stability and even-temperedness with negative emotionality, such as feeling anxious; And Openness to Experience describes the breadth, depth, originality and complexity of an individual’s mental and experiential life. The common method to measure these dimensions is the use of questionnaires such as the Neuroticism, Extraversion and Openness Five Factor Inventory (NEO-FFI) [22], and the Big-Five Marker scale (BFMS) [20].

Mood refers to baseline levels of affect that define peoples experiences. It is commonly modeled using the two dimensions called Positive Affect (PA) and Negative Affect (NA) scales [23]. PA and NA are related to corresponding affective trait dimensions of positive and negative emotionality [25]. PA reflects the extent to which a person feels enthusiastic, active, and alert. In contrast, NA is a general dimension of subjective distress and unpleasant engagement that includes aversive mood states, including anger, contempt, disgust, guilt, fear, and nervousness. In order to measure these two dimensions (PA and NA), Watson et al [24] developed the Positive and Negative Affect Schedule (PANAS) that consist of two 10-item mood scales, these schedules have shown to be internally consistent, uncorrelated and stable over a 2-month time period.

In the following we make a review of important databases that study affect personality and mood.

### 2.2 Databases for Affective Computing

Databases for the study of affective computing have been developed to allow researchers to compare results and reduce model development and testing times. In this section we review the most important publicly available databases for affect and personality research. We will review databases based on the modalities of video, neurological signals, and/or physiological signals. There is not, as far as we know, a single database developed for mood research.

Databases for the study of affect recognition based on video have focused mainly on the analysis of facial expressions. One of the main examples is the Sustained Emotionally Colored Machine-human Interaction using Nonverbal Expression (SEMAINE) database [8]. It consists of high-quality, multimodal recordings of 150 participants in emotionally colored conversations in a sensitive artificial listener (SAL) configuration, annotated for valence, arousal, and Facial Action Coding System (FACS) action units (AU). Another recent example is the Affectiva-MIT Facial Expression Dataset (AM-FED) [25], which consists of a comprehensively labeled dataset of ecologically valid spontaneous facial responses recorded in natural settings on the Internet. The dataset consists of 242 facial videos, frame-by-frame labels for the presence of 10 symmetrical and 4 asymmetrical AUs, 2 head movements, smile, general expressiveness, feature tracker fails, gender, location of 22 automatically detected landmark points and self-report responses of familiarity, liking, and desire to watch again. The database can be used for accurate AU detection on naturalistic and spontaneous data. The Denver Intensity of Spontaneous Facial Action (DISFA) database presented in [26] consists of well-labeled stereo video recordings of 27 adults while watching a 4-minute video clip. Labels consists of presence, absence, and intensity of 12 facial action units manually coded for each frame.

Databases for affect research based on physiological signals are also available. For instance, the MAHNOB-HCI [7] is a multimodal database that consists of synchronized recordings of face videos, audio signals, eye gaze data and peripheral/central nervous system physiological signals (ECC, GSR, respiration amplitude (RA), skin temperature (ST), and EEG) of 27 participants while watching first 20 videos, and second, short videos with relevant/non-relevant tags. Self-reports of the felt emotions using arousal, valence, dominance, and predictability scales, as well as emotional keywords, and agreement or disagreement with the tags are included. Koelstra et al present the DEAP database [1], with the purpose of implicit affective tagging from EEG and peripheral physiological signals (GSR, RA, ST, ECC, blood volume, Zygomaticus and Trapezius muscles Electromyogram (EMG), and Electrooculogram (EOG)) research. It consists of signals recordings of 32 participants while watching 40 music video clips. Self-assessment of arousal, valence, like/dislike, dominance and familiarity were obtained using Self-Assessment Manikins [27], and thumbs-up/thumbs-down symbols. For 22 participants, frontal face video was also recorded. A similar database that uses Magnetoencephalogram (MEG) is the DECAF database, which includes recordings of 30 participants to 40 one-minute music video and 36 movie clips. More recently, Zhang et al [6] collected the Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. It is an annotated, multimodal, multi-dimensional, spontaneous corpus of 140 participants from various ethnic origins, including 10 different emotion elicitation methods for specific target emotions (ee.g. surprise, disgust, fear). Recorded signals are 3D dynamic imaging, high-resolution 2D video, thermal sensing, electrical conductivity of the skin, respiration, blood pressure, and hearth rate. Facial video was annotated for the occurrence and intensity of facial AUs. These databases for affect research have not considered studying participants in group setting.

One of the first databases for personality research using video modality, is the Mission Survival II corpus [12]. It is a multimodal annotated collection of video and audio recordings (using 4 cameras and 17 microphones) of four meetings,
of 4 participants engaging in a mission survival task, in lab setting. Participants were annotated in terms of the Ten Item Personality Inventory [25] by 30 volunteers, to account for their personality states. These do not reflect the participants’ personalities but rather, moments where participants act more or less introvert/extravert, creative, etc. This dataset considers groups of participants, but it is not intended for affect research. A recent multi-modal database for implicit personality and affect recognition is the ASCERTAIN [29]. It includes recordings of the EEG, ECG, GSR, and facial video of 58 users, while viewing affective movie clips. They have shown that emotion-personality relationship is better captured by non-linear rather than linear statistics, and that personality differences are better revealed while comparing user responses to emotionally homogeneous videos. But this database only includes participants in individual configuration and does not share data about mood of participants. To the best of our knowledge there are not databases for personality research based on neurological or physiological signals and that studies participants in both individual and group configuration.

In Table 1 we summarize the characteristics of the reviewed databases and compare them to ours.

3 EXPERIMENTAL SETUP

In this section, we describe the experimental scenarios. We then make a description of the process we followed for the selection of stimuli, and we describe the modalities and equipment used in our experiments. Then, we give a detailed description of the experimental protocol. Finally, we describe the procedures we follow for affect self-assessment (internal annotation) and for external affect annotation (performed off-line by 3 annotators), and for participants’ personality and mood assessment.

3.1 Experimental scenarios

The main objective of this work is to study the personality, mood and affective responses of people in two social contexts, (i) when they are alone engaging with affective multimedia content (individual configuration), and (ii) when they are part of an audience and engage with the content at the same time with other people (group configuration). Additionally, we study people’s affective response to two types of eliciting content. The first type consists of short emotional videos (duration < 250s) selected to elicit specific affective states in the participants. The second type consists of long videos (duration > 14min), that present situations that could elicit various affective states over the participants, and where the story and the narrative could give context to the affective responses. As a result, we have designed two experiments, in the first one (Short videos experiment), all participants engaged individually with short affective videos. In the second experiment (Long videos experiment), the same participants engaged with long videos, but this time some of them did it individually, while the other ones did it in groups of four people.

Now we will proceed to describe the selection of the stimuli for both experiments.

3.2 Stimuli selection

Emotion elicitation depends greatly on a careful selection of the stimuli, which needs to be suitable for the objective of the study and allow for consistent results among trials [1]. In this work, we selected two sets of videos for emotion elicitation. The first one consists of short emotional videos and the second one consists of long videos. For the first set, 72 volunteers annotated, in the valence and arousal dimensions, the set of 36 emotional videos used in [1]. We then classified each of the videos into one of four quadrants of the valence-arousal (VA) space, namely high-valence/high-arousal (HVHA), high-valence/low-arousal (HVL), low-valence/high-arousal (LVHA), and low-valence/low-arousal (LVLA). From each quadrant, we selected the three videos that lay further to the origin of the scale, totaling 12 videos. Additionally, from the videos used in [7], we selected four videos, each corresponding to one of the four quadrants. The total number of selected short videos is 16, 4 for each quadrant of the VA space. We have preserved the IDs used in the original datasets. The selected short videos (51-150s long, $\mu = 86.7, \sigma = 27.8$) with their corresponding category on the VA space, and their IDs are listed in Table 2.

For the second set of videos, we initially selected 8 video extracts from movies based on their score in the IMDb Top Rated Movies list [30]. We selected movies that could allow us to extract a long segment ($\approx 20$min) which could be self-contained, did not require previous knowledge from the participants to be understood, and with strongly affective multimedia content (good combination of music and colors [31]). Four researchers watched them, review them, and tag them as belonging to one or more quadrants of the VA space. Finally, 4 videos were selected favoring the extracts that could evoke emotions in different quadrants of the VA space, and making sure all the quadrants were covered. The selected long videos (14.1-23.58min, $\mu = 20.0, \sigma = 4.5$) with their corresponding video ID, source and duration are listed in Table 3.

3.3 Neuro-Physiological Signals and Instruments

We recorded three main neural and peripheral physiological signals namely Electroencephalogram (EEG), Electrocardiogram (ECG), and Galvanic Skin Response (GSR). These modalities have shown good performance on affect estimation studies [32–34]. Below we give an introduction to each of them.

EEG: Electroencephalogram is the recording of electrical activity along the scalp. It measures voltage fluctuations resulting from ionic current flows within the neurons of the brain [35]. Cognitive processes have shown to be related to affect [17, 36], therefore EEG signals carry valuable information about the person’s affective state.

GSR: Galvanic skin response, also known as electrodermal activity (EDA), is a physiological signal that measures the electrical conductance of the skin [37]. It is measured through one or two sensors usually attached to some part of the hand or foot [38]. Skin conductivity varies with changes in skin moisture level (sweating) and can reveal changes in sympathetic nervous system related to the arousal level of a person as argued by Lang et al [34]. The changes in GSR are
TABLE 1
Summary of characteristics of databases for affect and personality. Last row is our database.

| Database       | No. Part. | Individual vs. Group | Purpose                                 | Modalities                                             | Annotations                               |
|----------------|-----------|-----------------------|-----------------------------------------|--------------------------------------------------------|------------------------------------------|
| DMAINE         | 150       | Individual            | Emotion recognition based on facial expressions | Audio and Visual                                      | Valence, arousal, and FACS.             |
| AM-FED         | 242       | Individual            | Spontaneous facial expression recognition “In-the-Wild” | Visual                                                 | 14 AUs                                   |
| DISFA          | 27        | Individual            | Spontaneous facial action recognition   | Visual                                                 | 12 AUs                                   |
| MAHNOB-HCI     | 27        | Individual            | Emotion recognition and implicit tagging | Visual, Audio, Eye Gaze, ECG, GSR, Respiration Amplitude, Skin temperature, EEG | Self-assessment of valence, dominance, predictability, and emotional keywords, Agreement/disagreement with tags. |
| DEAP           | 32        | Individual            | Implicit affective tagging from EEG and peripheral physiologival signals | EEG, GSR, Respiration Amplitude, Skin Temperature, Blood Volume, Electrocardiogram, and Electrooculogram. Visual for 22 participants. | Self-assessment of arousal, valence, liking, dominance and familiarity. |
| DECAF          | 30        | Individual            | Affect recognition                      | MI6, Near-infra-red facial video, horizontal Electrooculogram (hEOG), ECG, and trapezius-Electrooculogram (hEMG). | Self-assessment of arousal, valence, and dominance. Continuous annotation of valence and arousal of the stimuli. |
| Zhang et al corpus | 140     | Individual            | Emotional behaviour research            | 3D dynamic imaging, Visual, Thermal sensing, EDA, Respiration, Blood Pressure, and Heart Rate | Personality states by the Ten Item Personality Inventory. |
| Mission Survival II | 16      | 4 people group        | Personality states research             | Audio and Visual                                        | Personality states by the Ten Item Personality Inventory. |
| ASCERTAIN      | 38        | Individual            | Personality and Affect                  | EEG, ECG, GSR, and Visual                              | Big-Five personality traits, self-assessment of valence and arousal. |
| AMIGOS         | 40        | Individual & 4 people group | Affect, personality, mood and social context recognition | Audio, Visual, Depth, EEG, GSR, and ECG | Big-Five personality traits and PANAS. Self-assessment of valence, arousal, dominance, liking, familiarity, and basic emotions. External annotation of valence and arousal. |

TABLE 2
The Short Video Clips Listed With Their Sources (Video IDs are stated in parentheses). In the category column, H, L, A, and V stand for high, low, arousal and valence respectively.

| Category       | Excerpt’s source                                      |
|----------------|--------------------------------------------------------|
| HAHV           | Airplane (4), When Harry Met Sally (5), Hot Shots (9), Love Actually (80) |
| LAHV           | August Rush (10), Love Actually (13), House of Flying Daggers (18), Mr Beans’ Holiday (58) |
| LAVL           | Exorcist (19), My girl (20), My Bodyguard (23), The Thin Red Line (138) |
| HALV           | Silent Hill (30), Prestige (31), Pink Flamingos (34), Black Swan (36) |

TABLE 3
Selected Long Videos with Their ID, Source (Movie title. Director. Producer company. Released Year.), and Excerpt Duration.

| ID  | Source                                      | Duration |
|-----|---------------------------------------------|----------|
| N1  | The Descent. Dir. Neil Marshall. Lionsgate. 2005. | 23:35:0  |
| P1  | Back to School. Mr. Bean. Dir. John Birkin. Tiger Aspect Productions. 1994. | 18:43:0  |
| B1  | The Dark Knight. Dir. Christopher Nolan. Warner Bros. 2008. | 23:30:0  |
| U1  | Up. Drs. Pete Docter and Bob Peterson. Walt Disney Pictures and Pixar Animation Studios. 2009. | 14:06:0  |

In previous databases, neuro-physiological signals have been recorded using laboratory equipment (e.g. Biosemi ActiveTwo) which is expensive and limits the mobility of the participants. For this database we opted for recording the neuro-physiological signals using commercial wearable sensors that allow more freedom given that they use wireless technology. EGG was recorded using Emotiv EPOC Neuroheadset (14 channel, 128 Hz, 14 bit resolution). EGG channels according to the 10-20 [33] system are: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. ECG signal was recorded using the Shimmer 2R platform extended with an ECG module board (256 Hz, 12 bit resolution). ECG signal was recorded using three electrodes attached to the participant’s body. Two of them were placed at the right and left arm crooks and the third one at the internal face of the left ankle as reference. This set-up allows precise identification of heart beats as well as the full ECG QRS complex. GSR signal was recorded using the Shimmer 2R platform extended with a GSR module board (128 Hz, 12 bit resolution), with two electrodes positioned at the middle phalanges of the left hand’s middle and index fingers.

3.4 Video Recordings
Frontal face video was recorded in HD quality using a JVC GY-HM150E camera, positioned just below the screen. Additionally, depth and full body video was recorded using Microsoft’s Kinect V1 placed at the top of the screen. In this study we have not used the visual modality for prediction, but Mou et al [40], [41] have already shown the utility of our dataset using the visual information for prediction of affect, social context and group belonging. A participant during the...
short videos experiment and a group of participants during the long videos experiment can be observed in Fig. 1

3.5 Synchronization and Stimuli Display Platform

One PC (Intel Core i7, 3.4 GHz) was used to (i) present the stimuli, (ii) get and synchronize signals, and, in the case of the short videos experiment, (iii) obtain the self-assessment of participants. Shimmer sensors were paired to the PC using the Bluetooth standard, while the Emotiv headset paired using a proprietary wireless standard.

Videos were presented in a 40-inch screen (1280×1024, 60 Hz), each of them was displayed preserving the original aspect ratio and covering the highest screen-area possible. The remaining area was filled with black background. Subjects were seated approximately 2 meter from the screen. Stereo speakers were used and the sound volume was set at a relatively loud level, however it was adjusted when necessary.

3.6 Short Videos Experiment Protocol

Experiments were performed in a laboratory environment with controlled illumination. 40 healthy participants (13 female), aged between 21 and 40 (mean age 28.3), took part in the experiment. Prior to the recording session, each participant read and signed a consent form, and then read a sheet with instructions about the experiment. An experimenter was present there to answer any questions. When the instructions were clear to the participant, he/she was led into the experiment room. After that, the experimenter explained the meaning of the different affective scales used in the experiment and how to fill in the self-assessment form (See 3.8.1). Following, the sensors were placed and their signals checked with a test recording to assess the quality of the signal. Finally, the experimenter left the room, and the recording session began.

In order to assess the affective state of the participants before any stimuli had been shown, they were asked to perform an initial self-assessment for arousal, valence, and dominance, and selection of basic emotions (Neutral, Happiness, Sadness, Surprise, Fear, Anger, and Disgust). After that, 16 videos were presented in a random order in 16 trials, each consisting of: (1) A 5 second baseline recording showing a fixation cross. (2) The display of a small video. (3) Self-assessment of arousal, valence, dominance, liking, and familiarity, and selection of basic emotions (See 3.8.1). After the 16 trials, the recording session ended.

3.7 Long Videos Experiment Protocol

Subjects that participated in the short videos experiment, participated in the long videos experiment in either individual or group configuration. In the individual configuration, participants engaged in the experiment alone. In the group configuration, participants engaged in the experiment together with 3 other participants. Only 37 participants took part in the long videos experiment (participants 8, 24 and 28 were not available). 17 of them took part in this experiment in individual configuration, and 20 in group configuration (5 groups of 4 people). In order to maximize interactions, groups were formed to include people that knew each other, being either friends, colleagues, or people with similar cultural background. The IDs of participants that were in the individual configuration and in each group of the group configuration are listed in Table 4.

During the recording sessions, the participant(s) was(were) led to the recording room. While the different sensors were set up, experimenters explained the differences of the protocol compared to the short videos experiment. Every participant was given a set of self-assessment paper forms (See 3.8.1) and a pen, that were used to assess their affective state at the beginning and at the end of each video. Experimenters avoided to mention whether the participants could talk during the experiment, in order not to suppress or impose interactions. Once the sensors had been tested, the experimenters left the room and the recording session started.

The experiment consisted of displaying 4 long videos in random order. Videos were shown in two recording sub-sessions, each consisting of: (1) an initial self-assessment (45s) for arousal, valence, dominance, and selection of basic emotions. (2) the display, in two trials, of two long videos, each followed by (3) a self-assessment (45s) for arousal, valence, dominance, liking, familiarity, and selection of basic emotions (See 3.8.1). After the first sub-session a break of 15 minutes was given for the participants to rest, and during this time, they were offered refreshments. After the break, sensors’ signals were checked and the second recording sub-session started, after which the experiment was terminated.

After the long videos experiment, participants were asked to fill in, during their free time but as soon as possible, on-line forms with Personality Traits and PANAS questionnaires (See 3.9). Participants took 2 days on average to fill in the forms. Participants were given mugs and university gadgets in return for their participation once they filled in all required forms.

3.8 Affective Annotation

Self-assessment (internal annotation) refers to the process of directly asking a participant to rate her/his affective state while performing a task. It has the advantage of being the easiest, and possibly, the most direct way to assess the affective state of a person. At the same time, it is an intrusive process, subjects could be unreliable at reporting their emotions or they could hide their real emotions. Implicit assessment (external annotation) is a process that intends to assess the affective state of a person without her/him being actively involved in the process. The assessment is performed by external means like observing the person’s behavior, or by decoding the state by analyzing his/her physiological responses. For our experiments we opted
to perform both internal and external annotation to assess the affective state of participants. Details of the annotation processes are given below.

3.8.1 Participant’s Affect Self-assessment

At the beginning of the recording session of the short videos experiment, and of each of the two recording sub-sessions of the long videos experiment, participants performed a self-assessment of their levels of arousal, valence, and dominance, and were asked to select basic emotions that described what they were feeling at the start of each session/sub-session. Then, at the end of each trial, participants performed a self-assessment of the same levels and basic emotions as the initial self-assessment as well as liking, and familiarity that described what they felt during each video. Liking and familiarity were only assessed after the presentation of a stimulus.

The self-assessment form used for the short videos experiment can be seen in Fig. 2. Self-assessment manikins (SAM) [27] were used to visualize the scales of valence, arousal and dominance. For the liking scale, thumbs down/thumbs up symbols were used. This inquires the participants’ tastes, not feelings. The fifth scale asks the participants to rate their familiarity with the video. Arousal scale ranges from “very calm” (1) to “very excited” (9). Valence from “very negative” (1) to “very positive” (9). Dominance from “overwhelmed with emotions” (1) to “in full control of emotions” (9). The fourth scale ranges from disliking (1) to liking (9) the video. The familiarity scale ranges from “Never seen it before” (1) to “Know the video very well” (9). A continuous slider was placed at the bottom of each scale. Participants moved the slider to specify their self-assessment level. They were informed they could move the slider anywhere directly below or in-between of the manikins. Finally, participants were asked to select at least one of the basic emotions (Neutral, Disgust, Happiness, Surprise, Anger, Fear, and Sadness [16]), or as many as they felt during the video (a participant can consider a video to be both surprising and sad).

For the long videos experiment self-assessment, having a digital form for every participant of the groups was not practical, therefore we opted to use a paper version of the form in Fig. 2 for every participant of both individual and group configurations, in order to keep consistent the self-assessment between configurations.

In total, for the short videos experiment 17 annotations were obtained from each participant (1 at the beginning of the experiment, and 1 after each of the 16 short videos), and 6 annotations in the case of the long videos experiment (1 at beginning of the first recording sub-session, 1 after each of the two long videos of the first recording sub-session, 1 at the beginning of the second recording sub-session just after the 15 minute break, and 1 after each of the two long videos of the second recording sub-session). It is important to note that this annotation gives information related only to the participants’ initial and final affective states, not for specific instants during the videos.

3.8.2 External Affect Annotation

In order to provide a dataset that allows the study of the temporal evolution of affect, we off-line annotated, in the valence and arousal dimensions, the frontal videos of each participant recorded during the display of the stimuli of both experiments as follows.

First, the videos of a given participant recorded during the display of each of the 20 stimuli videos (16 short and 4 long), were manually cropped in order to show only a squared region around the face, covering from the top of the head to the start of the shoulders. We then split each of the participants’ face videos into 20 second clips. For this, the first 20 seconds of each video, including 5 seconds prior to the presentation of the stimuli, were extracted as first clip, then, starting from the 5s of the video (instant in which the stimuli started), \( n = \left\lfloor \frac{D}{(20s)} \right\rfloor \) non overlapping segments of 20s were extracted, with \( D \) being the duration
of the stimuli video in seconds. Finally, the last 20 seconds of the video were extracted as final clip. For every participant, \{6, 7, 5, 6, 4, 5, 8, 5, 7, 5, 9, 5, 4, 6, 7, 72, 58, 72, and 44\} clips were obtained respectively from videos \{4, 5, 9, 10, 13, 18, 19, 20, 23, 30, 31, 34, 36, 58, 80, 138, N1, P1, B1, and U1\}, totaling 340 clips per participant, 94 corresponding to the short videos and 246 to the long videos.

Three annotators rated the different clips (340 clips of 37 participants, totaling 12580 clips) in the valence and arousal dimensions. Both scales were continuous and ranged from \(-1\) (low valence/arousal) to \(1\) (high valence/arousal). The 20\(s\) clips from both experiments of one participant at a time were shown and annotated in a random order. Clips of a given participant were shown in the same random order to all annotators, however, the order of the clips was different for each participant. Since samples of the short and long videos experiments were randomly shown to the annotators, labels of the two experiments are directly comparable. The pipeline of the annotation consisted of the display of a randomly selected clip followed by the annotation performed by the annotator, first, of valence and then of arousal. Then, another clip was displayed and similar annotation was performed. This was done until all clips of a participant were annotated.

3.9 Personality and PANAS Assessment

The Big-Five personality traits were measured with an online form of the big-five marker scale (BFMS) questionnaire \[20\], in which, for each personality trait, using the basic question “I see myself as a person:”, ten descriptive adjectives are rated with a 7-point-likert-scale \[16\] and a mean is calculated.

The positive affect (PA) and negative affect (NA) schedules (PANAS) \[47\], were measured using the general PANAS questionnaire \[47\] that consists of two 10 questions forms, each to access the PA and the NA respectively. Participants filled in an on-line form rating their general feelings in a 5-point intensity scale using questions like “Do you feel in general...?” (e.g. active, afraid See \[47\]). PANAS is calculated by summing the values (between 1 and 5) of all 10 questions for PA and NA respectively, resulting in values between 10 and 50.

The distribution of the Big-Five personality traits, and PA and NA, over (i) the 37 participants that engaged in the long videos experiment, (ii) the 17 participants that engaged in the long videos experiment alone, and (iii) the 20 participants that did it in groups, are presented in Figure 3. Note that PA and NA scores have been scaled by a 0.1 factor. We can see that the distribution of ratings of the full population of participants, in comparison with the distribution from the participants of the individual configuration and the ones of the group configuration are very similar. The difference of distribution of ratings, for each of the seven dimensions of personality and PANAS, between the group and individual participants, is not significant \((p > 0.1)\) according to a two sample t-test for every dimension.

4 Data Analysis

In this section, we present a detailed analysis of the data gathered in both experiments.

4.1 Self-Assessment vs External Annotation

In order to validate the external annotations, we assessed the inter-annotator agreement. For this, we first mapped the annotations corresponding to each participant performed by every annotator to the \([0, 1]\) range, where 0 corresponds to low valence/arousal and 1 to high valence/arousal, and we calculated the Cronbach’s \(\alpha\) \[48\] statistic among annotators, commonly used for agreement assessment on continuous scales \[3\]. We obtained mean Cronbach’s \(\alpha\) over all participants of 0.98 for valence and 0.96 for arousal indicating a very strong inter annotator reliability for both dimensions.

With the objective to test at what degree, the affective state of participants assessed through self-assessment, is represented by the external annotations, we performed a comparison between the participants’ self-assessment and the external annotations of valence and arousal, for the short videos experiment. For each participant, we calculated the Spearman correlation coefficient as well as the \(p\)-value for the positive correlation test between the self-assessment scores of each video and the mean external annotation over all the annotators and all the segments of each video. Assuming independence, the resulting \(p\)-values were combined to one \(p\)-value using Fisher’s method \[49\]. We found that, for valence, the mean correlation over all participants is 0.44\((p < .05)\), and 0.15\((p < .05)\) for arousal. These correlations are statistically significant which indicates that the external annotation is a good predictor of the affective state of participants, though for the arousal dimension the correlation is low which shows that it is easier to externally assess valence than arousal.

In Figure 4(a), the distribution of the self-assessment of valence and arousal of all participants for the short videos experiments (16 samples per participant) can be observed. Annotations of all participants have been mapped to the \([-1, 1]\) range. The graph includes circles representing the mean scores, over all participant, of each video. We can observe in this graph that in general valence elicitation worked better than arousal, showing a well defined separation between low and high valence stimuli. Even though the separation of arousal is not as prominent, still there is a difference between low and high arousal stimuli. Figure 4(b) shows the distribution of the external annotation of valence and arousal over the 16 videos of the short videos experiment (94 samples by participant). The mean scores, over all

![Fig. 3. Distribution of the Big-Five Personality Traits (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness) and Positive Affect and Negative Affect Schedules (PA and NA) for (i) All, (ii) Individual configuration, and (iii) Group configuration participants of the Long Videos Experiments. PA and NA are scaled by a 0.1 factor.](image-url)
the 20-second clips of each video and over all participants are marked with circles. In this graph we observe that the data shows a V-shape relating valence and arousal, which is a result of the difficulty of eliciting high-levels of arousal with neutral valence, and high/low levels of valence with low arousal. We also can observe that in general participants showed the expected affective states (e.g. participants showed higher valence/arousal) with high-valence(arousal) content in comparison to low-valence(arousal) content), though the difference is not as clear as in self-assessment (fig. 4(a)).

4.2 Analysis of Valence and Arousal for Individual and Group Configurations

The external annotations of both experiments have been analyzed to test if valence and arousal, expressed by the participants, differed depending on whether they were alone or in a group. We have considered two sets of participants separately. The first set (individual set) corresponds to the 17 participants that were alone in the long videos experiments, and the second set (group set) corresponds to the 20 participants that did it in groups. In Fig. 5 we show the differences in annotations for valence and arousal for the individual set in comparison with the group set in both short and long videos experiments. Fig. 5(a) and (d) show the mean valence and arousal annotations for (i) individual participants (red curve), (ii) group participants (blue curve), and (iii) all participants (black dashed curve), for each of the 340 20s clips. The clips are shown by the video they are part of, and ordered according their appearance in the video. To show differences of affect expressed by individuals in comparison to groups, in the figure, clips where the difference in the distribution of scores for the group set are significantly lower or higher (p < 0.05 according to a two sample t-test) than the distribution of scores of the individual set are marked with black points, and have been shadowed (orange for group scores < individual scores, and gray for group scores > individual scores). Fig. 5(b) and (e), show the mean annotations of valence and arousal, for the same sets of participants, of the clips of the short videos experiment, while Fig. 5(c) and (f) present the mean annotations for the clips of the long videos experiment. In the (b), (c), (e), and (f) graphs, samples are ordered according to the mean score over all participants (dashed black curve). The clips for which the difference between the distribution of scores from individuals and groups is significant (p < 0.05 according to a two sample t-test) are marked with black points.

We can see from Fig. 5(a) and (d), that both the high and low areas of the valence and arousal dimensions are covered between all the videos. Comparing the graphs of the short videos experiment (Fig. 5(b) and (e)) with the ones of the long videos experiment (Fig. 5(c) and (f)), we can see that, for the short videos experiment where all participants were alone, 21.3% of the clips present significant differences in valence between group and individual participants, and are concentrated in the low valence region, and 21.1% of the clips present significant differences in arousal. In the case of the long videos experiment, where some participants were in groups, in valence, 25.6% of the clips present significant differences between groups and individuals, and it is important to note that 48% of the clips with significant differences appear in the high valence region (mean valence > 0). For arousal, 26.4% of the clips present significant differences between groups and individuals. Another interesting behaviour can be observed in Fig. 5(f), where it is easy to note that in the long videos experiment group participants showed lower levels of arousal for clips with low arousal clips and higher levels of arousal for high arousal segments. These show that the group configuration has an important effect on the valence and arousal expressed by the participants.

We calculated the Spearman correlation coefficient ρ and the p-value between the social context label and the mean external annotations for valence and arousal of clips of the long videos experiments. The social context label was considered 0 if the participant was alone in the long videos experiment and 1 if her/him was in a group. We found that, there is a significant positive correlation (ρ = 0.37, p < 0.05) between the social context and the mean valence. This significant correlations imply that, in the long videos experiment, participants that were in groups showed higher valence than participants that were alone. We did not find significant correlation between the social context and the arousal scores (p > 0.05), which suggest that the social context does not have a common effect in the arousal expressed by the participants for all clips.

Additionally, we tested whether the scores for valence and arousal were different depending on whether the participants were alone or accompanied in the long videos experiment. From the graphs in Fig. 5(c) and (f) we noted that the scores for the clips with lower levels of valence/arousal, present a different behavior than the ones of higher valence/arousal. Because of this, we have performed the analysis for the low and high valence/arousal independently. For each of the two cases (valence of long videos, and arousal of long videos), we sorted the clips in increasing order of affective score, then we classified the half of clips with the lower scores as low class (e.g. low-valence) and the half with the higher scores as high class (e.g. high-valence). We performed a two samples t-test of the mean scores of valence/arousal between the individual and
In this section we present baseline methods and results from neuro-physiological signals. We first describe the extracted features for the used social context for the long videos experiment, and the mean external annotations of valence and arousal, first, for all graphs, clips where the distribution of scores of individual participants is significantly different than the one of group participants ($p < 0.05$) according to a two sample t-test, are marked with black points. In the case of (a) and (d), video IDs are indicated in the captions. Clips where the distribution of scores of individual participants is significantly higher than the one of group participants ($p < 0.05$), are highlighted in orange. Clips where the distribution of scores of group participants is significantly higher than the one of individual participants are highlighted in gray. In the case of (b), (c), (d), and (f) the horizontal axis represent the number of clips. Origin of valence and arousal (horizontal axis at ($V = 0$)) and (A = 0)) divides the scale into high-valence (HV: $V > 0$) and low-valence (LV: $V < 0$), and into high-arousal (HA: $A > 0$) and low-arousal (LA: $A < 0$).

4.3 Affect, Personality, Mood and Social Context Correlations

In Table 5 we show the Spearman inter-correlation that is observed between the dimensions of personality, PANAS, and social context for the long videos experiment, and the mean external annotations of valence and arousal. In the case of the correlations of personality and PANAS dimensions with external annotations of valence and arousal, first, for every participant we have obtained the mean value of the external annotations of valence and arousal for the samples of (i) the short videos experiment, and (ii) the long videos experiment.

In the case of personality and PANAS, we got positive significant correlations ($p < 0.05$) between extraversion and agreeableness, agreeableness and both conscientiousness and PA, and conscientiousness and emotional stability. We also observed that NA is mostly negatively correlated to all personality and PA dimensions. In the case of social context, we did not get significant differences in personality and PANAS distribution between individual and group participants, which imply that the group participant and the individual participant have common distribution of personalities (e.g. individual participants have similar levels of extraversion than group participants). In general, correlations between personality and PANAS with respect to valence and arousal were not significant, which implies that personality and mood does not necessarily affect the levels of valence and arousal expressed by the participants, with the exception of PA which showed significant positive correlation (0.61) with respect to the mean value of arousal of the long videos, which indicates that high-PA participants showed higher levels of arousal (they were more active) than the low-PA participants.

### Table 5

| Dims. | Ag. | Co. | E.S. | Op. | PA | NA | S.C. |
|-------|-----|-----|------|-----|----|----|------|
|       |     |     |      |     |    |    |      |
| Ex.   | 0.44* | 0.09 | 0.21 | 0.13 | 0.32 | -0.48* | 0.20 | 0.01 | 0.02 | 0.05 | 0.18 |
| Ag.   | -    | 0.34* | 0.14 | 0.24 | 0.43* | -0.41* | 0.18 | 0.21 | 0.60 | 0.13 | 0.21 |
| Co.   | -    | 0.35* | 0.01 | 0.26 | -0.26 | 0.07 | 0.12 | 0.14 | 0.13 | 0.19 |
| E.S.  | -    | -    | 0.24 | -0.12 | -0.44* | 0.03 | 0.21 | 0.11 | 0.18 | 0.15 |
| Op.   | -    | -    | -    | 0.20 | -0.35* | 0.04 | 0.23 | 0.13 | 0.06 | 0.02 |
| PA    | -    | -    | -    | -    | -0.03 | 0.03 | 0.16 | 0.30 | 0.41* |
| NA    | -    | -    | -    | -    | -    | -0.01 | -0.28 | -0.02 | 0.12 | 0.04 |
|       |     |     |      |     |    |    |      |

In this section we present baseline methods and results for prediction of affect (valence and arousal), personality, PANAS and social context using neuro-physiological signals. We first describe the extracted features for the used
modalities, then we present our method for single modality and fusion of modalities for single-trial classification of affect. After that, we present our method for single-trial classification of personality traits, PANAS and social context, using single modalities and different schemes for fusion of modalities. Finally, we present and discuss our results.

5.1 EEG, ECG, and GSR Features

We used the neuro-physiological modalities of EEG, ECG and GSR, to record the participants’ implicit responses to affective content. Now we will describe the extracted features from the employed modalities. All the features were calculated using the signals recorded during each of the 340 twenty-second clips described in section 3.8.2. Different to other studies that use the concatenation of ECG and GSR as one modality, we study each of them independently to account better for the contribution of each one to the recognition task. The summary of features is listed in Table 6.

EEG: Following [1], we extracted power spectral densities (PSD) features from the EEG signals. For this, the EEG data was processed using the sampling frequency of 128 Hz. The signals were average-referenced, and high-pass filtered with a 2 Hz cut-off frequency. We removed eye artefacts with a blind source separation technique [50]. By employing the Welch method with windows of 128 samples (1.0s), PSDs, between 3 and 47 Hz, of the signals of every clip were calculated for each of the 14 EEG channels. The obtained PSDs were then averaged over the frequency bands of theta (3-7 Hz), low alpha (8-10 Hz), alpha (8-13 Hz), beta (14-29 Hz), and gamma (30-47 Hz), and their logarithms were obtained as features. Additionally, the spectral power asymmetry between the 7 pairs of symmetrical electrodes, in the five bands, was calculated. We got 14 channel * 5 bands, and 7 symmetrical channels * 5 bands, totaling 105 PSD features for every sample (See Table 6).

ECG: Using the method reported in [51], we accurately localized the heart beats in ECG signals (R-peaks) to calculate the inter beat intervals (IBI). Using IBI values, we calculated the heart rate (HR) and heart rate variability (HRV) time series. Following [7] and [51] we extracted 77 features (See Table 6).

GSR: Following the state of the art method of Kim [51], we calculated the skin conductance (SC) from GSR and then normalized the SC signal. We low-pass filtered the normalized signal with 0.2 Hz and 0.08 Hz cut-off frequencies to get the low pass (LP) and very low pass (VLP) signals, respectively. Then, we de-trended the filtered signals by removing the continuous piecewise linear trend in the two signals. We calculated 31 GSR features employed in [1], [7] (See table 6).

5.2 Single Trial Classification of Affect from Short and Long Videos

For single trial affect (valence and arousal) classification, we first mapped the features of every modality for each recording session to the [-1, 1] range in order to avoid the baseline differences that are natural to different recording sessions. This was done for every participant, considering each of the 4 long videos as a recording session and the session of the 16 videos of the short videos experiment as a fifth session. For each of the modalities (EEG, ECG, and GSR), we have tested three scenarios, the first one considers to train and test the system only with the samples of the short videos experiment (94 samples by participant). The second considers only the samples of the long videos experiment (246 samples by participant), and the third one considers the combination of the samples of all the videos of both experiments (340 samples by participant), giving in total 9 recognition tasks for every affect dimension. We use a leave-one-participant-out cross validation in which, in order to predict each affect dimension d label, for each participant i a Gaussian (G) Naïve Bayes (NB) classifier is trained. A NB G assumes independence of the features and is given by: $G(f_1, \ldots, f_n) = \arg\max
\p_{C=c} \prod_{i=1}^{n} p(F_i = f_i | C = c)$ where F is the set of features and C the classes. $p(F_i = f_i | C = c)$ is estimated by assuming Gaussian distributions of the features and modeling these from the training set. In each step of the cross validation, from the N available participants, the samples of one participant are used as the test set, and the samples of the remaining N-1 participants are used as the training set. For feature selection, Fisher’s linear discriminant J [52] defined as $J(f) = \frac{|\mu_1 - \mu_0|}{\sigma_1^2 + \sigma_0^2}$, is calculated for each feature from the training samples. Features are then sorted in decreasing order according to their J value, and with a second 10-fold cross-validation over the training set, the optimal [1 : h] most discriminative features are selected. Then, the classifier is trained over all the samples of the training set using the selected features, and it is tested in the test set.

For each of the three scenarios (short, long and all videos), we also explored feature level fusion of the single modality features, in which, previous to feature selection, we concatenated all the features of the three modalities, and then performed feature selection and trained the classifier in the same way as for the single modalities.
5.3 Classification of Personality, PANAS and Social Context on Short and Long Videos

5.3.1 Classification Methods

For personality traits and PANAS prediction, we associated the features of the samples of all the videos of a given scenario (short, long, and all videos), for each of the modalities (EEG, ECG, and GSR), to the label of the different personality and PANAS dimensions. Also, from the samples of the long videos and the samples of all the videos (short and long), we have tried to predict the social context of participants (whether a participant was alone or in a group in the long videos experiment). For every participant, 8 support vector machine (SVM) classifiers with radial basis function (RBF) kernel were trained, one for each of the 5 personality traits, 2 for mood dimensions of PA and NA, and 1 for social context prediction. First, for every participant, the features have been mapped to the range in the same way as in the affect recognition experiment. In order to reduce dimensionality, we averaged the features of all the samples of the videos of a given scenario (e.g., 94 samples of 105 features for the short videos scenario and the EEG modality) into one sample consisting of the mean values of the features of every modality for every participant. We use a leave-one-participant-out cross-validation. During training, principal components analysis (PCA) is performed in the training set for reducing further the dimensionality to 35 components for the EEG and ECG features given the number of training samples (36), and 31 for the GSR features, and the trained PCA is applied to the test set. Using a second leave-one-participant-out cross-validation over the training set, we selected the optimal combination of the PCA component, C parameter and sigma parameter for the SVM RBF, using grid-search of the C and sigma, in combination with each of the 35 main PCA channels. With the selected PCA channel, C and sigma, the final SVM with RBF kernel is trained over all the samples of the training set and tested on the test set. Train and test labels are divided into high and low classes using the median value of each personality and mood dimension as threshold. Social context recognition was not performed in the short videos experiment because every participant was alone in that experiment.

5.3.2 Fusion of Modalities

In order to use complementary information of the different modalities, we have implemented both, feature level fusion (early integration), and decision level fusion (late integration) of the three modalities (EEG, ECG and GSR) for prediction of personality traits, PANAS and social context. We implemented the next methodologies for feature and decision level fusion:

Feature Level Fusion (FLF): In FLF we trained the PCA independently on the features of each modality and then concatenated the PCA components of all modalities, into a feature vector of dimensionality $35 \times 2 + 31 = 101$. The classifier is trained in the same way as the single modality, though the search is over the 101 components.

Mean Decision Level Fusion (MeanDLF): In MeanDLF, we calculated the mean predicted probability over the three modalities for the low-class and the high-class. If the mean probability for high-class was greater than the probability for low-class the sample was labeled as high-class, and low-class otherwise.

Maximum Decision Level Fusion (MaxDLF): For MaxDLF, maximum predicted probability was used, in which we considered only the modality with highest probability associated to low-class and high-class for the given sample. If the maximum probability was for high-class, the sample was labeled as high-class, and low-class otherwise.

Optimal Decision Level Fusion (OptDLF): For OptDLF, we use the linear combination model given by: $p_0 = \sum_{i=1}^{n} \alpha_ip_i$, where $n$ is the index of modalities ($1 \rightarrow \text{EEG}, 2 \rightarrow \text{GSR}, 3 \rightarrow \text{ECG}$), $0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^{n} \alpha_i = 1$. We trained the $\alpha_i$s on the probabilities predicted from the three modalities using a linear regressor on training set, and then test them on the test set. After combination, if $p_0 > 0.5$ the sample was labeled as high-class, and low-class otherwise. Note that MeanDLF is an instance of the model where $\alpha_i = 1/n$ for $i = 1, 2, ..., n$.

5.4 Results and Discussion

In Table 7, the average F1-scores (average F1-score for both classes), over all participants, for classification of valence and arousal, with the different modalities using the Gaussian Naïve Bayes classifier, are presented. Three scenarios are included, the first considers only the samples of the short videos experiment, the second the samples of the long videos experiment, and the third one the samples of all the videos of both experiments. Results for feature level fusion of the three modalities are also included. The table also includes the random baseline results (analytically determined) that are obtained by assigning labels randomly.

Random levels for all the scenarios for valence and arousal had 0.5 mean F1-score each. We observe significant higher than chance ($p < .01$ according to an independent one-sample t-test) F1-scores for all the scenarios using the EEG modality, for the long videos and all videos scenarios using ECG, and only for arousal recognition in the long videos and all videos scenarios using GSR. In general, arousal identification got higher performance than valence, except for ECG modality in the long videos experiment. For all scenarios of valence and arousal recognition, EEG got significantly higher performance than ECG and GSR ($p < 0.0001$ for both), resulting in a mean improvement, over the three scenarios, of $2.2\%$ and $3.2\%$ for recognition of valence and arousal over the ECG, the second best performing modality. ECG is still significantly better ($p < 0.05$) than
the GSR modality. Feature level fusion does not outperform the best performing modality but it is still significantly higher than chance ($p < 0.01$). Prediction of valence and arousal for short videos outperforms the predictions for the long videos but the differences are not significant ($p = 0.32$ for valence and $p = 0.19$ for arousal). Using the videos of both experiments for training and testing does not increase the performance for recognition of valence and arousal in comparison to training and testing only in the short or in long videos. Our baseline results show average performance compared with the ones reported in the literature for similar experiments of valence and arousal recognition 11, 11, 11.

In Table 8, the mean F1-score of the positive and negative classes over all participants for binary classification of personality traits, PANAS, and social context is presented. Three scenarios are shown, the first using only the samples of the short videos experiment, the second using the samples of the long videos experiment, and the third using all the samples from both experiments. We have also implemented the baseline method proposed by Abadi et al. [13], based on a linear regression model for prediction using two physiological modalities, namely EEG and physiological signals (ECG+GSR). We applied their method in the same 37 participants that we use in this study, but only in the short videos experiment as they did, for the sake of comparison. The empirically estimated baseline results obtained by randomly assigning the labels are also reported.

Random

| Videos | Modality | Evt | Agr | Con | Emot | Open | Pa. | Na | So. Co |
|--------|----------|-----|-----|-----|------|------|-----|----|--------|
| Short  | EEG      | 0.86 | 0.87 | 0.82 | 0.85 | 0.62 | 0.75 |    |        |
|        | GSR      | 0.86 | 0.86 | 0.82 | 0.85 | 0.62 | 0.75 |    |        |
|        | ECG      | 0.40 | 0.69 | 0.69 | 0.64 | 0.64 | 0.67 |    |        |
| Long   | EEG      | 0.85 | 0.85 | 0.82 | 0.85 | 0.62 | 0.75 |    |        |
|        | GSR      | 0.85 | 0.85 | 0.82 | 0.85 | 0.62 | 0.75 |    |        |
|        | ECG      | 0.40 | 0.69 | 0.69 | 0.64 | 0.64 | 0.67 |    |        |
| All    | EEG      | 0.72 | 0.75 | 0.86 | 0.86 | 0.56 | 0.67 | 0.70 | 0.75 |
|        | GSR      | 0.38 | 0.59 | 0.62 | 0.55 | 0.67 | 0.70 | 0.58 | 0.65 |
|        | ECG      | 0.41 | 0.69 | 0.69 | 0.64 | 0.64 | 0.67 | 0.53 | 0.61 |

For recognition of social context, EEG from all videos outperformed the other modalities (0.858 F1-score). Interestingly, in the case of EEG and GSR, considering all the videos (both short and long) for prediction of social context outperforms using only long videos, which implies that the inclusion of the information embedded in the samples of the short videos, introduces important information for social context prediction.

If we consider recognition from all the dimensions, and all scenarios, in general EEG is the best performing modality (0.669 mean F1-score), the advantage is significant against both GSR ($p = 0.012$) and ECG ($p = 0.015$) modalities according to a two sample t-test. The best scenario for prediction of both personality traits and PANAS is the long video scenario, with a mean F1-score of 0.625 over all the dimensions and modalities, but the difference is not significant in comparison with the short videos, and all videos scenarios ($p = 0.193$ and $p = 0.395$ respectively). The best scenario for prediction of social context is the combination of all videos (0.708 mean F1-score), but the difference is not significant in comparison with the long videos scenario ($p = 0.367$).

Comparing our method with the baseline one presented in [13], if we consider only the short videos experiment, with EEG modality we manage to outperform the baseline study in prediction of all but openness dimension. Considering the physiological signals (ECG and GSR), our method, using the single modality of GSR, outperforms the baseline study in prediction of agreeableness, conscientiousness, openness and NA. ECG is better only for prediction of conscientiousness, emotional stability and NA. Considering the long videos scenario and the all videos scenario as well, our method outperforms the baseline one for all dimension. The main advantage of our method is that it can learn from non-linear relations between the features and the labels, whereas the method of [13] only learns linear relations.

Table 9 presents the mean F1-score over all participants for binary classification of personality traits, PANAS and social context, for the feature level fusion and the decision level fusion schemes described on 5.3.1. We present the same three scenarios as for the single modality experiments (short, long and all videos). We have included the results of
the best performing modality for each scenario for the sake of comparison.

If we consider only the short videos scenario, feature level fusion using FLF outperformed the best single modality for prediction of emotional stability and NA, and matched it in conscientiousness prediction. DLF using mean probability (MeanDLF) only outperformed the single modality for prediction of conscientiousness and openness. DLF using the maximum probability (MaxDLF) only outperforms slightly the best single modality in prediction of openness. For the long videos experiment the single modality is only outperformed in prediction of PA using FLF, and in prediction of agreeableness and emotional stability using MaxDLF. When using the samples of all the videos, only DLF using the optimal combination (OptDLF) for prediction of agreeableness outperforms the best single modality. The best fusion schema is FLF with a mean F1-score of 0.633 over all dimensions and scenarios, but it still falls short in comparison with EEG, the best performing single modality (See Table 9).

To summarize Tables 9 and 10, the best performance for prediction of extraversion is obtained using long videos with EEG modality (0.782 F1-score). For agreeableness, it is OptDLF using all the videos (0.757 F1-score). For conscientiousness MeanDLF, using the short videos, gives the best performance (0.888 F1-score). For emotional stability, EEG using all videos is the best (0.865 F1-score). For openness, MaxDLF using the short and all videos, match the GSR modality using all the videos (0.674 F1-score). For PA, EEG modality using all the videos gets the best performance (0.770 F1-score). And for NA prediction, the best performance was for FLF using the short videos (0.854 F1-score).

6 Conclusions

In this work we presented AMIGOS, a database for research of affect, personality traits, mood, affect, and social context by means of neuro-physiological signals. We elicited affect with both short and long videos in both individual viewers and group of viewers configuration. We found significant correlations between internal and external affect annotations of valence and arousal, indicating that external annotation is a good predictor of the affective state of participants. We showed that when participants are in groups they tend to express more positive reactions than when they are alone, which reflects as an increase in the mean valence for group participants in comparison with individual participants. EEG was the best modality for prediction of valence and arousal for the short videos, long videos and combination of short videos and long videos experiments, while feature level fusion did not improve the results. For single-trial classification of personality traits, PANAS, and social context we obtained significant results, where EEG was the best modality, and the combination of short and long videos yield the best performances. Finally, feature level fusion improved the results for NA prediction, whilst decision level fusion did it for prediction of agreeableness and conscientiousness. The database is being made publicly available for other researchers to try their methods of affect, personality, mood, and group vs. individual configuration research.

References

[1] S. Koelstra, C. Muehl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “DEAP: A database for emotion analysis using physiological signals,” Affective Computing, IEEE Transactions on, vol. 3, no. 1, pp. 18–31, 2012.

[2] G. Chanel, C. Rebetez, M. Brancourt, and T. Pun, “Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty.” IEEE Transactions on Systems, Man, and Cybernetics, Part A, vol. 41, no. 6, pp. 1052–1063, 2011.

[3] B. O’Neill, Toward A Computational Model of Affective Responses to Stories for Augmenting Narrative Generation. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 256–263.

[4] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, “DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses,” IEEE Transactions on Affective Computing, vol. 6, no. 3, pp. 209–222, July 2015.

[5] S. Koelstra, C. Muehl, and I. Patras, “Eeg analysis for implicit tagging of video data,” in Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, 2009, pp. 1–6.

[6] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang et al., “Multimodal spontaneous emotion corpus for human behavior analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3438–3446.

[7] S. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A Multi-modal Database for Affect Recognition and Implicit Tagging.” T. Affective Computing, vol. 3, no. 1, pp. 42–55, 2012.

[8] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversation between a Person and a Limited Agent,” IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 5–17, 2012.

[9] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, “The more the merrier: Analysing the affect of a group of people in images,” in Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013), Ljubljana, Slovenia, J.-S. L. Ramzan, R. van Zwal, Ed., 2015.

[10] M. Wilson, “Mrc psycholinguistic database: Machine-readable dictionary, version 2.00,” Behavior Research Methods, Instruments, & Computers, vol. 20, no. 1, pp. 6–10, 1988.

[11] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, “Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines.” American Psychologist, vol. 70, no. 6, pp. 543–556, Sep. 2015.

[12] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, “A multimodal annotated corpus of consensus decision making meetings,” Language Resources and Evaluation, vol. 41, no. 3, pp. 409–429, 2007.
