EMOEEG: a New Multimodal Dataset for Dynamic EEG-based Emotion Recognition with Audiovisual Elicitation

Anne-Claire Conneau*, Ayoub Hajlaoui†*, Mohamed Chetouani† and Slim Essid*

* LTCI, Télécom ParisTech, Université Paris–Saclay

†Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie

Abstract—**EMOEEG is a multimodal dataset where physiological responses to both visual and audiovisual stimuli were recorded, along with videos of the subjects, with a view to developing affective computing systems, especially automatic emotion recognition systems. The experimental setup involves various physiological sensors, among which electroencephalographic sensors. The experiment is performed with 8 participants, 4 from both genders. The stimuli include both sequences of static images from the IAPS dataset, and short video excerpts focusing on negative fear-type emotions. The annotation is obtained by participant self assessment, after a calibration phase. In the case of video stimuli, a novel simplified dynamic annotation strategy is used to enhance the quality and consistency of the self-assessments. This paper also analyses the annotation results and provides a statistical study of inter-annotator agreement. The dataset will continue to grow and will be made publicly available.**

Index Terms—**Electroencephalography (EEG), Multimodal Data, Affective Computing, Fear-type Emotions, Valence, Arousal, Annotation, Inter-annotator agreement**

I. INTRODUCTION

Having a better grasp on physiological manifestations of human emotion would be beneficial for affective computing research. Contributions to automatic emotion recognition mainly rely on modalities such as speech, facial expressions, or eye gaze. The main limitation of these modalities is their alterability, whether voluntary or not [1]. In contrast, physiological modalities such as electroencephalography (EEG) do not suffer from such a drawback. Compared to other physiological modalities, EEG has the advantage of capturing information related to internal emotional states which may not be reflected externally. Thus, EEG has attracted the attention of researchers in the field of affective computing and it has been shown to hold precious cues for emotion classification [2].

In order to achieve such a goal, synchronized multimodal datasets are needed, with enough experimental repetitions for each subject to capture intra-subject variability of physiological signals. Indeed, the individuality of physiological responses plays a major role, whether it be at the feature extraction level or at the stage of assessing the emotion [3]. Also, to be applied in real-world scenarii, multimodal and dynamic stimuli should be employed, but only a few corpora fulfill these requirements.

In this paper, the focus is put on a complex emotion recognition setting for which multimodality is interesting, along the line of the MAHNOB-HCI initiative [4]. The latter provides multimodal recordings of subjects who watched audiovisual stimuli for emotion recognition and implicit tagging research purposes. So do the eNTER-FACE'06 [5] and DEAP [6] data collection initiatives. However, these datasets have some limitations, notably the fact they do not take into account the dynamics of emotional states, that is their variation over time during the exposition to such stimuli.

With this in mind, we have performed synchronized multimodal recordings of subjects while visual and audiovisual stimuli were presented to them. The visual stimuli were extracted from the IAPS [7] (International Affective Picture System) database, whereas the audiovisual stimuli were extracted from the SAFE corpus [8]. Self-annotation was made dynamically, each participant giving both global and variational feedback (dynamic annotation) on each stimulus after being exposed to it. The originality of our dataset lies in three main aspects:

- Repetitions were performed on a per subject basis for the purpose of a reliable intra-subject classification. This accommodates the known fact that the brain activity is subject specific, especially as far as physiological manifestations of affective states are concerned.
- A calibration phase was designed, using specific images so that a participant could get familiar with the valence-arousal space, at the beginning of the process, then be able to refer to those calibration stimuli during the self-annotation that follows each subsequent image or video stimulation.
- Moreover, a novel dynamic annotation scheme was implemented, for the case of video stimuli, that adopts a simplified strategy in order to make the produced self-annotations more robust, and favor their con-

sistency (both on an intra-subject and inter-subject basis).

The description of the protocol and the signals recorded in our database is made in Section II, followed by the description of the emotion self-assessment strategy in Section III. Section IV discusses inter-annotator statistics. Finally, we give our conclusions in Section V.

II. APPARATUS

A. Stimuli choice

EEG responses have a high inter-subject variability, as shown for instance by the results that Lee and al. obtained on the IDIAP database [3]. Therefore, being able to perform effective EEG-based emotion recognition usually requires numerous recording instances per subject. In the case of video stimuli, the duration of each stimulus has to be long enough (at least 10 seconds) to allow dynamic emotion elicitation. On the other hand, the duration of the experiment should remain reasonably short in order to control the cognitive load on the subject. Too long experiments would indeed result in an unsatisfactory concentration level throughout the session. Consequently, we chose to limit the duration of a session to 90 minutes, which was verified to be reasonable through some preliminary tests. This has constrained the number of images and video stimuli that we chose to exploit, as described below.

1) Visual stimuli: The images were extracted from the IAPS database. The IAPS database is composed of images which were self-assessed by 100 annotators [7]. Following the approach chosen by the eNTERFACE'06 initiative [5], the visual stimuli were assembled as blocks of 5 related images each. We actually used a selection of blocks considered in [5] in order to facilitate the comparison with this dataset. A total of 50 blocks corresponding to 250 images was exploited. Each block belongs to one of these 3 classes: neutral (average valence, low arousal), positively excited (high valence, high arousal) and negatively excited (low valence, high arousal). Figure 1 represents the distribution of the selected images in the valence/arousal space [9], [10], where we can clearly distinguish 3 clusters corresponding to those 3 classes.

2) Audiovisual stimuli: In order to observe emotion dynamics, the video stimuli were chosen to be slightly longer than the blocks of image stimuli. Most videos were selected from the SAFE corpus, in addition to 6 videos related to phobias and 2 emotionally neutral videos. This choice is motivated by the development of strategies amenable to the analysis of the impact of violent videos on humans, and possibly treatments for subjects suffering from phobia. Thus, in terms of valence and arousal, there is a bias towards negative emotions in the choice of video stimuli. The SAFE corpus contains 401 excerpts extracted from 30 movies. Each excerpt is divided into segments of variable
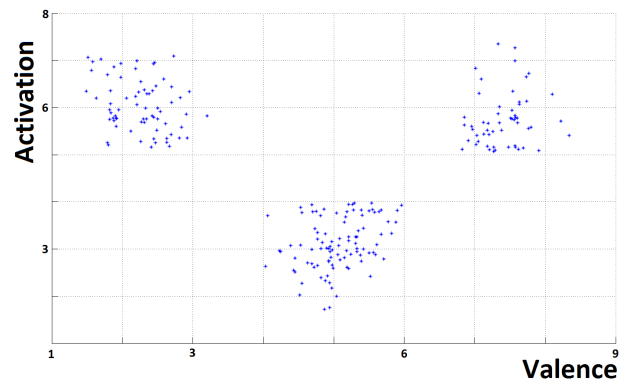

Fig. 1. Distribution of extracted images in the valence/arousal space

length corresponding to a variation of elicited emotion, with a focus on negative fear-type emotions. Excerpts from this corpus which best show the emotion dynamics were selected according to the standard deviation of user annotations characterizing the evolution of the imminence and intensity of danger in each video excerpt. A total of 100 video stimuli was used. Each video lasts approximatively 15 seconds, and each session is composed of 50 videos. The sound volume was normalized for all videos, in order to avoid arousal bias that could result from loud audio in some videos.

B. Recording equipment and synchronization

A wireless B-Alert X24[1] headset recorded the EEG, EOG (Electrooculogram), EMG (Electromyogram) and ECG (Electrocardiogram) of the participant, whereas an Affectiva bracelet[2] recorded both skin conductance and temperature. A PC with a 64-bit operating system was used to play the stimuli, record all the signals and manage the protocol procedure, whereas a full HD TV (165cm) played the stimuli. The physiological recordings were timestamped by the PC. The Affectiva bracelet was synchronized with the system timestamping of the PC before the beginning of the experiment. Further, a discrete HD camera was placed in front of each participant, as well as a second HD control camera which allowed the experimenter, located in another room, to check the attention level of the participant. A microphone was directed towards the participant. A Linear Time Code (LTC) was recorded in one of the audio channels of the HD video cameras in order to accurately locate its frames across time. This LTC signal was initialized at the beginning of each experiment.

C. Experimental protocol

Each session was composed of a calibration phase (Section III-C), after which 25 eNTERFACE'06 blocks were

[1]http://www.advancedbrainmonitoring.com/xseries/x24/
[2]http://qsensor-support.affectiva.com/

randomly presented to the participant, followed by 50 short videos. To take into account the persistence of an emotional state after the end of the stimulus playback, the subject looked at a white screen for 10 seconds, right after watching a block of images or a video (Figure 2.). After that, he/she assessed his/her global valence and arousal levels during the stimulation (global or static annotation), with discrete values ranging between 1 (very negative) and 9 (very positive). In the case of audiovisual stimuli, the subject dynamically annotated the stimuli .

D. Summary

Table I summarizes the EMOEEG database properties. A total of eight subjects participated to the experiment. Three of them went through two different sessions, where different stimuli were used. We use this to evaluate the ability of emotion recognition systems to generalize across different sessions. This makes a total of 11 sessions. All participants signed an informed consent.

The dataset will be made publicly available through a companion website[3]. Also, the recordings will continue in the next fews months and new recordings will be posted as they become available.

III. RECORDED SIGNALS AND ANNOTATION

A. EEG and other physiological signals

EEG signals were recorded at a sampling frequency $f_s = 256$ Hz. The 20 EEG electrodes of the B-Alert X24 headset that we used follow the 10-20 international system. ECG, EOG, and EMG were recorded by the same system, with other electrodes. EEG signals are usually polluted by artifacts resulting from various sources, such as ocular or muscular movement. However, rather than being considered as mere artifacts, the corresponding additional ECG, EOG and EMG signals can also be used for emotion recognition, as shown in [11], [12]. In addition to these

---

[3]http://www.tsi.telecom-paristech.fr/aao/en/2017/03/03/emoeeg-a-new-multimodal-dataset-for-dynamic-eeg-based-emotion-recognition-with-audiovisual-elicitation/

Fig. 2. Protocol for one stimulus

TABLE I
SUBJECTS AND CORRESPONDING SESSIONS

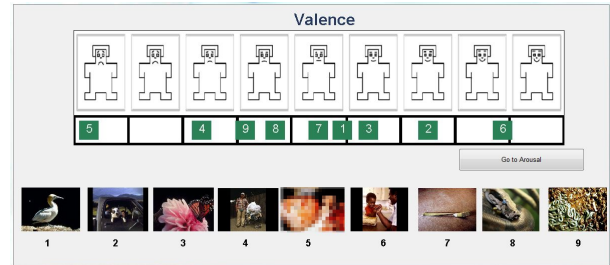|	
Number of participants	8, 5 male and 3 female
Number of sessions	11
Participants who took 2 sessions	5,6,8
Number of stimuli per session	25 image blocks, 50 videos
Duration of an image repetition	25,5s
Duration of a video repetition	28s
Physiological signals recorded	EEG, EOG, EMG, ECG, EDA

Fig. 3. Example of a ranking of calibration stimuli in terms of valence.

signals, skin conductance and temperature were recorded. They could be used both for emotion classification or as a way of controlling the consistency of emotional state self-annotation.

B. Emotion self-assessment strategy

The subjects were asked to evaluate the emotion they felt and not the emotion they might attribute to the stimulus (sometimes called perceived emotion). For instance, if the video excerpt came from a horror movie, but did not frighten them, they were not expected to annotate fear. They were asked to locate the emotion they felt in the valence/arousal space [10].The annotation was made using a modified version of the Self Assessment Manikin (SAM) [9] approach. The modification consists in displaying, along with the manikins, the images annotated in the calibration phase (see below) so they can serve as references. It is indeed acknowledged that it is easier for an annotator to proceed in a relative fashion, comparing the valence/arousal values of a stimulus to be assessed to some references (here the calibration images), as opposed to rating them in an absolute fashion [13].

C. Calibration phase

One of the important characteristics of EMOEEG is the introduction of a calibration phase performed by each participant in order to (i) become familiar with the notions of valence and arousal, and (ii) to identify his/her limits on each dimension. The calibration phase was composed of 9 images selected from the reference IAPS database. These images were chosen to span over the whole extent of the valence/arousal space and correspond to low standard deviation values in terms of valences and arousals reported by IAPS annotators (high reference annotation confidence). The participant was asked to annotate and sort these 9 images using a discrete annotation scale ranging from 1 to 9 for both the valence and arousal dimensions. Figure 3 depicts the annotation process during the calibration. The subject placed each of the images on the valence or arousal axis appearing below the manikins, where each green icon corresponds to a different image, identified by its number.

D. Global and dynamic annotation

Global annotation allows participants to evaluate the emotion elicited by each stimulus and it is performed after each block of images selected from the eNTERFACE'06 protocol. It was also performed for each audiovisual stimulus.

Dynamic annotation allows them to describe potential changes of affective states during a stimulus. Our approach differs from the one proposed by Gtrace[4], which requires that the participant annotate each dimension dynamically in a synchronized way with the video. In that case, he/she has to watch the entire video a second time, after its initial playback, the one that is considered to elicit the emotion to be analyzed. This implies longer sessions and may cause exhaustion in the sense that watching the video and annotating it synchronously induces a supplementary cognitive load. Finally, video playback repetitions may cause habituation and affect the capacity of the used to annotate the elicited emotion, the one felt during the first stimulus playback.

To alleviate these issues, we propose a new dynamic emotion annotation strategy, less costly both in time and in cognitive load, in addition to an improvement regarding habituation problems. Our dynamic annotation gathers both dimensions (valence and arousal) in one window, so that the participant can annotate them simultaneously, which subsequently reduces the duration of the annotation phase. Moreover, our configuration does not imply annotating while watching the video. Rather, the participant watches the video only once, after which he/she has to remember the variation of the emotion he/she felt while watching it. The variation is re-transcribed only roughly as a three-segment annotation, namely three values per dimension, respectively corresponding to the beginning, the middle and the end of the video. No specific time division is imposed to the participant, that is the exact location of beginning, middle and end segments is not to be specified. Figure 4 illustrates this annotation approach. Subjects who watched the same video may thus choose different values of valence (or arousal) that would still exhibit the same dynamic trend (e.g. completely flat, increasing then decreasing, etc.).

E. Inter-annotator agreement statistics.

To evaluate the quality of the self-assessments made by the subjects who took part in our recordings, we analyzed the annotation reliability by computing various inter-annotator statistics. In this part, we report inter-annotator agreement analysis with the IAPS reference labels (performed by the IAPS annotators) of visual stimuli and traditional statistics (Cohen's kappa and correlation).

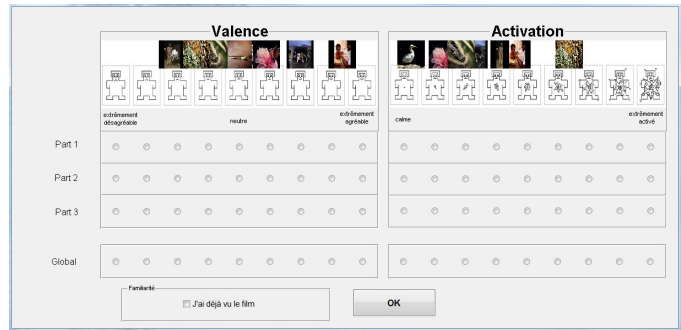[4]https://sites.google.com/site/roddycowie/work-resources

Fig. 4. Dynamic video self-annotation.

TABLE II
PEARSON'S CORRELATION COEFFICIENTS BETWEEN IAPS ANNOTATIONS AND EACH EMOEEG'S SUBJECT ANNOTATIONS ON ENTERFACE'06 IMAGE BLOCKS.

Subject	Valence correlation	Arousal correlation
1	0,69	0,48
2	0,95	0,81
3	0,88	0,54
4	0,88	0,49
5	0,91	0,74
6	0,96	0,77
7	0,97	0,56
8	0,93	0,76
Average	0,90	0,64

1) Correspondence between EMOEEG and IAPS annotations: We computed Pearson's correlation coefficients between each subject's annotations and the IAPS reference annotations. We consider it as a measure of consistency between the emotion each annotator felt for a given stimulus and what that stimulus generally elicits. Table II reports these coefficients for each subject and for each dimension. It reveals stronger agreement for the valence dimension. For both dimensions, it is made clear that correlation coefficients are subject-sensitive.

2) Inter-annotator agreement within EMOEEG: Of the 50 image blocks and 100 video excerpts used as stimuli in the corpus, all the subjects did not watch the same ones. Therefore, the number of stimuli that are common to a pair of subjects varies from 9 to 50 for visual stimuli and 22 to 50 for audiovisual ones.

Cohen's kappa coefficient κ [14] measures the agreement between two annotators. Table III indicates the mean of kappa coefficients among all pairs of subjects for different types of stimuli. For each pair, the coefficient is computed over the stimuli that the two subjects had in common. A value higher than 0.61 is considered as good, and between 0.41 and 0.60, it is considered as moderate. As expected, the table shows that higher agreements are obtained with image stimuli and static annotations.

As shown in Table IV, in the case of audiovisual stimuli, the use of dynamic annotations improves inter-annotator

TABLE III
MEAN OF KAPPA COEFFICIENTS AMONG ALL PAIRS OF SUBJECTS FOR DIFFERENT TYPES OF STIMULI

Stimuli	Valence	arousal
Images	0.75	0.30
Videos (global+dyn. annotation)	0.18	0.08
Images and videos (global ann.)	0.49	0.18

TABLE IV
MEAN OF KAPPA COEFFICIENTS AMONG ALL PAIRS OF SUBJECTS FOR BOTH TYPES OF ANNOTATION IN THE CASE OF AUDIOVISUAL STIMULI

Annotation	Valence	Arousal
Global	0.12	0.08
Dynamic	0.19	0.08

agreement when it comes to valence, whereas arousal inter-annotator agreement is not impacted.

In affective computing, ranking emotional experiences is more relevant than quantifying them [13]. In such a context, Pearson's correlation might not be adequate: therefore, we also computed Spearman's correlation, which does not compute the correlation between the values of two variables, but rather between the ranks of such values. Table V reports the means of p-values obtained on all pairs of subjects using Pearson's or Spearman's correlation, both with image annotations and video static annotations. P-values represent the probability of wrongly assuming that the correlation is greater than 0. They are computed using a Student's t distribution in Pearson's case, whereas in Spearman's, exact permutation distributions are used. The results show that even if the Spearman's coefficient does not give better results in all cases, it allows the p-value to drop under 0.05 in the arousal case, for which Table II showed lower agreement than for valence.

Moreover, Table V shows better p-values for videos than for images in the arousal case, whether it be using Pearsons's or Spearman's coefficient. This could seem in contradiction with the results obtained using Cohen's kappa in Table III. However, the nature of such coefficients is different. In addition, the variation of arousal might be clearer from one video to another than from one block of images to another.

TABLE V
MEAN OF PEARSON AND SPEARMAN P-VALUES

Coefficient	Images	Videos (global)
Pearson (valence)	0.003	0.03
Spearman (valence)	0.01	0.03
Pearson (arousal)	0.07	0.03
Spearman (arousal)	0.06	0.02

IV. CONCLUSIONS

We have recorded a multimodal affective dataset at the disposal of the affective computing community. It contains various modality recordings such as physiological responses (EEG, ECG, EMG, EOG, skin conductance, temperature information) and videos of the subjects. It offers a significant number of experimental repetitions per subject, which is not common for this kind of task. It also comes with an innovative annotation strategy which both exploits a calibration phase and a dynamic approach of emotion elicitation. Finally, inter-annotator statistics have been presented and discussed. They show higher inter-annotator agreements for valence than for arousal. Further work will focus on the use of annotation dynamics for emotion recognition.

REFERENCES

[1] G. Fiebig and M. Kramer, "A framework for the study of emotions in organizational contexts," *Management Communication Quarterly*, vol. 11, no. 4, pp. 536–572, 1998.

[2] V. Bajaj and R. Pachori, "Human emotion classification from eeg signals using multiwavelet transform," in *Medical Biometrics, 2014 International Conference on.* IEEE, 2014, pp. 125–130.

[3] H. Lee, Y.-D. Kim, A. Cichocki, and S. Choi, "Nonnegative tensor factorization for continuous eeg classification," *International journal of neural systems*, vol. 17, no. 04, pp. 305–317, 2007.

[4] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[5] A. Savran, K. Ciftci, G. Chanel, J. Mota, L. Hong Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut, "Emotion detection in the loop from brain signals and facial images," 2006.

[6] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[7] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (iaps): Affective ratings of pictures and instruction manual," *Technical report A-8*, 2008.

[8] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, and C. Sedogbo, "The safe corpus: illustrating extreme emotions in dynamic situations," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006)). Genoa, Italy*, 2006, pp. 76–79.

[9] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[10] A. Mehrabian and J. Russell, *An approach to environmental psychology.* the MIT Press, 1974.

[11] E. André, L. Dybkjaer, W. Minker, and P. Heisterkamp, *Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings.* Springer Science & Business Media, 2004, vol. 3068.

[12] E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn, "Analysis of physiological signals for recognition of boredom, pain, and surprise emotions," *Journal of physiological anthropology*, vol. 34, no. 1, p. 25, 2015.

[13] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, 2014.

[14] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.