A Protocol for Cross-Validating Large Crowdsourced Data: The Case of the LIRIS-ACCEDE Affective Video Dataset

Yoann Baveye[1][2], Christel Chamaret[1]
[1]Technicolor
975, avenue des Champs Blancs
35576 Cesson Sévigné, France
yoann.baveye@technicolor.com
christel.chamaret@technicolor.com

Emmanuel Dellandréa[2], Liming Chen[2]
[2]Université de Lyon, CNRS
Ecole Centrale de Lyon, LIRIS
UMR5205, F-69134, Lyon, France
emmanuel.dellandrea@ec-lyon.fr
liming.chen@ec-lyon.fr

ABSTRACT

Recently, we released a large affective video dataset, namely LIRIS-ACCEDE, which was annotated through crowdsourcing along both induced valence and arousal axes using pairwise comparisons. In this paper, we design an annotation protocol which enables the scoring of induced affective feelings for cross-validating the annotations of the LIRIS-ACCEDE dataset and identifying any potential bias. We have collected in a controlled setup the ratings from 28 users on a subset of video clips carefully selected from the dataset by computing the inter-observer reliabilities on the crowdsourced data. On contrary to crowdsourced rankings gathered in unconstrained environments, users were asked to rate each video through the Self-Assessment Manikin tool. The significant correlation between crowdsourced rankings and controlled ratings validates the reliability of the dataset for future uses in affective video analysis and paves the way for the automatic generation of ratings over the whole dataset.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems—*Multimedia databases*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology, Video*

General Terms

Experimentation, Human Factors, Reliability

Keywords

Crowdsourced annotations, Experimental validation, Interrater reliability, Affective video datasets, Affective computing

1. INTRODUCTION

Mining the affective behavior of people when watching TV or movies is important for many real life applications (*e.g.*, personalized content delivery, affective video indexing, movie recommendation, video summarization and so on). However, it is often time-consuming to collect significant annotated data sources for machine learning. The recent years have witnessed crowdsourcing as a powerful tool to rapidly collect large amount of data, including in the field of affective computing [14, 20, 24]. Nevertheless, because of the open nature of crowdsourcing, data collected through such a process are potentially noisy due to a certain observer variability and mostly require further post-processing to validate the annotations gathered from various uncontrolled environments.

Recently, a large affective video dataset LIRIS-ACCEDE [2] annotated using crowdsourcing has been made available to the academic community[1]. In this dataset, the emotions elicited by the video clips are described in a 2D emotion space with two widely used continuous-valued emotion dimensions: valence and arousal. Valence can range from negative (*e.g.*, sad, disappointed) to positive (*e.g.*, happy, elated), whereas arousal ranges from inactive (*e.g.*, calm, bored) to active (*e.g.*, excited, alarmed). LIRIS-ACCEDE is composed of 9,800 video excerpts that last from 8 to 12 seconds, extracted from 160 movies shared under Creative Commons licenses. All the excerpts are ranked along the induced valence and arousal axes thanks to annotations gathered on a crowdsourcing platform. Trusted annotators were asked to compare two excerpts and select the one that induced the most positive emotion for valence and the calmest emotion for arousal. However, the pairwise annotations made in various uncontrolled environments have not been validated yet. That is why the contribution of this study is to create an experimental protocol where 28 participants are asked to rate in a controlled environment using the Self-Assessment-Manikin scales a subset of the dataset carefully selected by computing the Krippendorff's alpha coefficient. The significantly high correlation between crowdsourced rankings and controlled ratings cross-validates the dataset and finally allows to better understand its bias in the affective space.

The paper is organized as follows. An overview of the method is given in Section 2 while Section 3 presents how the film clips are selected from the dataset. Next, in Section 4, the experimental protocol of the user study is described. In an attempt to validate the dataset, the results are discussed in Section 5. Finally, conclusion and future work end the paper in Section 6.

[1] http://liris-accede.ec-lyon.fr/

CrowdMM'14, November 7, 2014, Orlando, FL, USA.
Copyright 2014 ACM 978-1-4503-3128-9/14/11 ...$15.00.
http://dx.doi.org/10.1145/2660114.2660115.

2. OVERVIEW

We collected more than one million annotations from 3,959 crowdworkers to rank in the 2D valence-arousal space the 9,800 video clips included in the LIRIS-ACCEDE dataset [2]. Even if unnoticeable test questions were randomly inserted throughout the tasks to reject untrustworthy crowdworkers, several unknown factors can potentially affect such a large number of annotations and require further post-processing to validate the data:

1. Crowdworkers were asked to focus on what they felt in response to the video excerpts but it is unknown whether they did so or not. Indeed, it is possible to make judgments on the basis of conventional characteristics without experiencing any emotion [23].

2. In the experiment described in [2], there was no way to make sure that crowdworkers really turned on the volume to judge the videos.

3. The 3,959 crowdworkers made the annotations in various uncontrolled environments under different conditions which may create significant noise in the data since the elicitation of an emotion is a subtle process that can be influenced by the physical and social environments [21].

That is why it is essential to cross-validate the crowdsourced annotations collected in [2] by designing a new experimental protocol in a controlled environment.

With the growing interest of researchers in crowdsourcing, the cross-validation between highly subjective annotations collected either using a crowdsourcing platform or a controlled environment is becoming a crucial step to legitimate the use of crowdsourcing in such tasks. Chen *et al.* used the Kendall τ distance to measure the consistency of the pairwise judgments for four case studies conducted in laboratory or on a crowdsourcing platform [6]. They showed that the use of crowdsourcing does not compromise the quality of the results and that the judgments provided by both experiments were reasonably consistent. Ribeiro *et al.* used a MOS-based testing methodology to demonstrate the consistency of the subjective quality ratings collected on a crowdsourcing platform [19]. Using the Blizzard Challenge 2009 dataset, they showed that the crowdsourced ratings are highly correlated with the ratings made by participants in a controlled environment. More recently, Redi *et al.* investigated how well recognizability and aesthetics ratings collected in a controlled lab environment can be replicated by a crowdsourcing environment [18]. They showed that the standard deviations of the scores assessed by the participants in both experiments were similar and that their MOS (Mean Opinion Scores) were positively correlated.

In all these works, the datasets used to demonstrate the consistency of the crowdsourced annotations are small enough to be completely annotated in laboratory by a small amount of participants. In this work, it is not conceivable to collect ratings in laboratory for all the 9,800 excerpts of the LIRIS-ACCEDE dataset. As a consequence, the first step of this work is to select a subset of excerpts from the dataset to create an experiment of acceptable duration. Furthermore, the distribution in the 2D valence-arousal space of the dataset used in this work is unknown because only rankings were collected on the crowdsourcing platform in [2]. To solve this limitation, discrete ratings are collected in the new controlled experiment presented in this paper, allowing us to understand the range of emotions elicited by the dataset.

3. SELECTING STIMULI FROM THE LIRIS-ACCEDE DATASET

Eliciting emotional reactions from test participants in laboratory experiments is quite tricky, that is why it is crucial to select most effective stimuli. Therefore, Krippendorff's alpha measure has been computed to select a subset of the dataset to be used in the user study. It ensures that the highest reliable film clips in eliciting induced emotions are selected. Krippendorff's alpha reliability coefficient is a generalization of several known reliability measures [10]. It applies to any number of observers, any number of categories, any metric, incomplete or missing data, and large or small sample sizes not requiring a minimum. The reliability of the excerpt $i \in \{0, \ldots, 9799\}$ for arousal or valence is defined as:

$$\alpha^i = 1 - \frac{D_0^i}{D_e^i} \tag{1}$$

where D_0^i is the observed disagreement among values assigned to pairwise comparisons in which one of the two compared excerpts is the excerpt i and D_e^i is the expected disagreement when the annotations are attributable to chance:

$$D_0^i = \frac{1}{n_i} \sum_c \sum_k o_{ck}^i \cdot \delta_{ck}^2 \tag{2}$$

$$D_e^i = \frac{1}{n_i\,(n_i - 1)} \sum_c \sum_k n_c^i \cdot n_k^i \cdot \delta_{ck}^2 \tag{3}$$

with c and k the categories of annotations' values, n_i the number of pairable annotations, *i.e.* $n_i = \sum_c \sum_k o_{ck}^i$, and n_c^i, n_k^i the number of pairable annotations with value c and k respectively, *i.e.* $n_c^i = \sum_k o_{ck}^i$ and $n_k^i = \sum_c o_{ck}^i$. The coincidence matrix o_{ck}^i is defined as:

$$o_{ck}^i = \sum_{u \in U^i} \frac{\text{Number of } c \text{ - } k \text{ pairs in comparison } u}{m_u - 1} \tag{4}$$

with U^i the subset of the pairwise comparisons for which one of the two compared excerpts is the excerpt i and m_u the number of annotations for comparison u.

Actually, the LIRIS-ACCEDE dataset has been annotated using forced-choice pairwise comparisons [2] so that each video excerpt is accompanied by two discrete values ranging from 0 to 9799 representing its arousal and valence ranks. In concrete terms, a comparison between two video clips was displayed to workers until three annotations were gathered, *i.e.* $m_u = 3, \forall u \in U^i, \forall i \in \{0, \ldots, 9799\}$. The crowdworkers had to select the excerpt that conveyed the most the given emotion in terms of valence or arousal. Thus, c and k values in eq. (4) represent the excerpt selected by a crowdworker and can be equal to "excerpt 1" or "excerpt 2". The δ^2 coefficient for nominal data is defined as:

$$\delta_{ck}^2 = \begin{cases} 0 & \text{if } c = k \\ 1 & \text{if } c \neq k \end{cases} \tag{5}$$

In other words, for each excerpt, eq. (1) is used to compute its Krippendorff's alpha coefficient for valence using the crowdsourced annotations of valence and its Krippendorff's alpha coefficient for arousal using the crowdsourced annotations of arousal. These values are used to select 20 excerpts per axis (valence and arousal) that are regularly distributed in order to get enough excerpts to represent the whole dataset in the 2D valence-arousal space while being relatively few to create an experiment of acceptable duration. For each axis, the 20 excerpts that form a perfect regular distribution are the ones so that their rank equals to $\frac{9800}{19} \times n$ with $n \in \{0, \ldots, 19\}$. These ranks are called the optimum ranks. Thus, for each axis and each optimum rank, we select the excerpt i with $\alpha^i \geq 0.6$ as close as possible to the optimum rank. This process ensures that the 40 selected film clips have different levels of valence and arousal and thus are representative of the full dataset. They are also representative of the agreement among the crowdworkers since just half of the video clips are highly reliable in eliciting valence or arousal, *i.e.* the 20 video clips that are highly reliable in eliciting valence may not be highly reliable in eliciting arousal and vice versa.

4. RATING OF SELECTED STIMULI

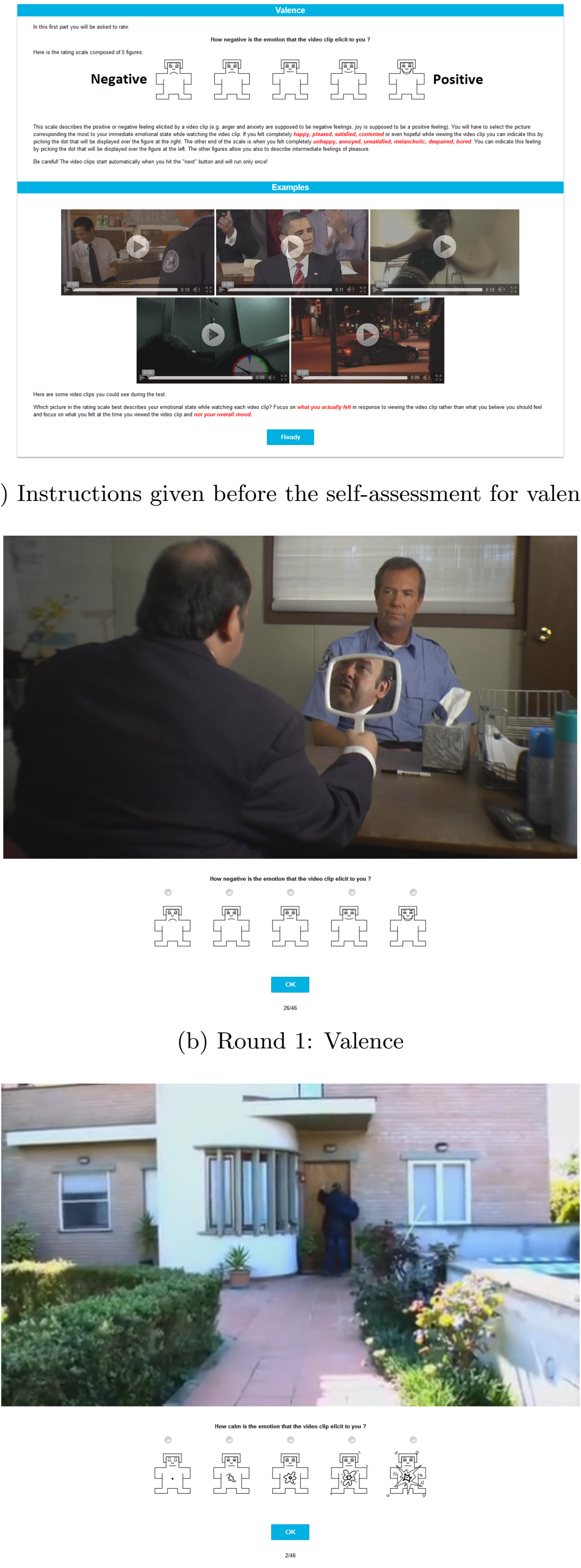
One of the objectives of this user study is to provide ratings of arousal and valence for each of the 40 film clips selected in the previous section.

4.1 Protocol

28 volunteers participated in the experiment (8 females and 20 males), aged between 20 and 52 ($mean = 34,93 \pm 8,99$). All the participants are working at Technicolor as researchers, PhD candidates or trainees. Of these individuals, 23 are French and the others are Bangladeshi, Chinese, Ethiopian, Romanian or Vietnamese. 8 out of the 28 participants participated in the experiment in the morning. They were asked to read a set of instructions informing them of the protocol of the experiment and the meaning of the two different scales used for self-assessments (see Fig. 1(a)). Following the procedure of Philippot [17], participants were instructed to report what they actually felt while watching the video excerpts, rather than what they thought they should feel. They were also asked to focus on what they felt at the time they watched the film clips, rather than their general mood of the day. Moreover, they were told that they were free to withdraw from the test at any time. Five test video clips from the LIRIS-ACCEDE dataset, but different from the 40 videos selected for this experiment in Section 3, were shown to the participants to make them understand the type of stimuli they could see during the test. An experimenter was also present at the beginning to answer any questions.

In addition to the 40 film clips selected in the previous section, 6 videos from these film clips were repeated twice in order to measure the intra-rater reliability. Consequently, 46 film clips (but 40 unique videos) were shown to the participants. The videos were presented in a dark room on a 22-inch screen (1,920 ×1,200, 60 Hz) in S-RGB mode and all film clips were displayed with a resolution height of 780px, the width depending on the ratio of each video. Participants were seated approximately 1 meter from the screen. The stereo Sennheiser PXC 360 BT headphone was used and the volume was set at a comfortable level.

The scale to report the ratings was the Self-Assessment Manikin (SAM) [4]. It is an efficient pictorial system to be used in experiments which utilizes sequences of humanoid figures to represent emotional valence, arousal and dominance. Due to its non-verbal design it can be used conveniently



In this first part you will be asked to rate:

How negative is the emotion that the video clip elicit to you ?

Here is the rating scale composed of 5 figures:

This scale describes the positive or negative feeling elicited by a video clip (e.g. anger and anxiety are supposed to be negative feelings, joy is supposed to be a positive feeling). You will have to select the picture corresponding the most to your immediate emotional state while watching the video clip. If you felt completely ***happy, pleased, satisfied, contented*** or even hopeful while viewing the video clip you can indicate this by picking the dot that will be displayed over the figure at the right. The other end of the scale is when you felt completely ***unhappy, annoyed, unsatisfied, melancholic, despaired, bored***. You can indicate this feeling by picking the dot that will be displayed over the figure at the left. The other figures allow you also to describe intermediate feelings of pleasure.

Be careful! The video clips start automatically when you hit the "next" button and will run only once!

Here are some video clips you could see during the test.

Which picture in the rating scale best describes your emotional state while watching each video clip? Focus on ***what you actually felt*** in response to viewing the video clip rather than what you believe you should feel and focus on what you felt at the time you viewed the video clip and ***not your overall mood.***

(a) Instructions given before the self-assessment for valence

How negative is the emotion that the video clip elicit to you ?

○ ○ ○ ○ ○

26/46

(b) Round 1: Valence

How calm is the emotion that the video clip elicit to you ?

○ ○ ○ ○ ○

2/46

(c) Round 2: Arousal

Figure 1: Screenshots of the interface used for the experiment.

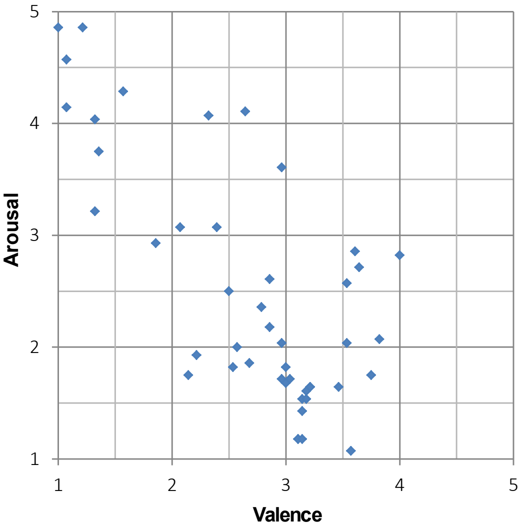
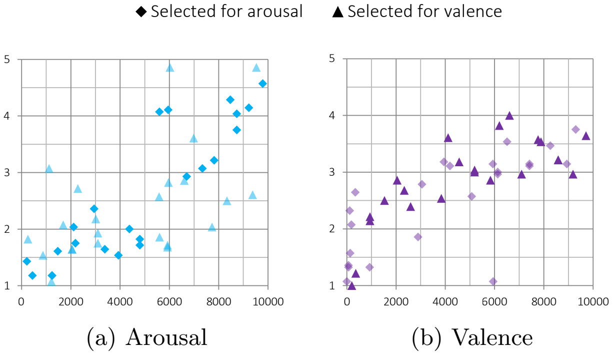


Figure 2: Distribution of the 46 film clips in the affective space (mean values for valence and arousal).



(a) Arousal

(b) Valence

Figure 3: Correlation between rankings (horizontal axis) and ratings (vertical axis) for both arousal and valence for the 46 films clips. A distinction is made between the 23 film clips selected for arousal and the others selected for valence.

regardless of the age or cultural background. Affective ratings were made on the discrete 5-point scale versions for valence and arousal. Instructions were adapted from Lang *et al.* [11] (see Fig. 1(a)).

The experiment took place in two rounds. During the first round, participants performed a self-assessment of their level of valence directly after viewing each film clip. All the videos were presented in a random order and the participant was asked to rate immediately "How negative is the emotion that the video clip elicits to you?" (see Fig. 1(b)). For the second round, participants performed a self-assessment of their level of arousal according to "How calm is the emotion that the video clip elicits to you?" (see Fig. 1(c)), all the videos being also presented in a random order. The video excerpts were run only once but participants had unlimited time to rate the videos. The next video started immediately once the participant hit the "OK" button. The vocabulary used in this test to describe the valence and arousal is the same than the one used to annotate the whole dataset in [2]. Valence has been intentionally annotated before arousal because it is intuitively easier to assess and thus more encouraging and motivating.

4.2 Results

Fig. 2 shows the distribution of the ratings of valence and arousal, suggesting that negative film clips were rated as more arousing than positive ones. This correlation is not surprising since Lang *et al.* showed that only specific areas of the 2D valence-arousal space are relevant [12]. The distribution displayed in Fig. 2 is also similar to those depicted in previous works dealing with the affective impact of multimedia content [8, 11, 13] , except that the distributions illustrated in these works show much more data eliciting positive and arousing emotions (see Section 5.2 for further discussion).

Globally, the mean standard deviation of the ratings is higher for arousal ($SD = 0.771$) than for valence ($SD = 0.631$), indicating that participants agreed more when assessing valence. It is confirmed by the Krippendorff's alpha, measuring the inter-annotator agreement, which is higher for the self-assessment of the level of valence ($\alpha = 0.282$) than for the self-assessment of their level of arousal ($\alpha = 0.225$). Both values are positive which indicates that there is an agreement between annotators despite the subjectivity of the experiment and are comparable to other studies dealing with affective computing [13, 15]. A two-factor (Women, Men) ANOVA failed to revealed significant gender differences. It is interesting to mention that another two-factor (AM, PM) ANOVA revealed that participants who started the experiment in the afternoon tend to report greater levels of arousal ($F = 30.1, p = 1.79 \times 10^{-6}$). This observation is consistent with the findings of Soleymani *et al.* indicating that average arousal ratings in response to videos increase with time of day [25].

The intra-rater reliability can also be computed thanks to the 6 film clips that have been annotated twice by each annotator. The mean-square error (MSE) of the ratings of the duplicated film clips is very low for valence ($MSE = 0.002$) as well as for arousal ($MSE = 0.021$) meaning that the repeatability of the experiment is high for a short period of time and consequently that annotators understood the scales and did not answer at random. To qualify these high intra-rater reliabilities, it is worth considering that due to the short duration of the experiment, some participants could have remembered the score given to the first occurrence of the video clip, thus reducing the impact of this criterion. Nevertheless, it is consistent to use these ratings to validate the crowdsourced annotations of the LIRIS-ACCEDE dataset.

5. CROSS-VALIDATION & BIAS OF THE LIRIS-ACCEDE DATASET

5.1 Cross-validation

The results from the controlled rating experiment presented in this paper allow to cross-validate the dataset that has been previously ranked in the affective space thanks to numerous crowdsourced pairwise annotations gathered in [2] from various uncontrolled environments. They also allow to better understand the distribution and the bias of the dataset in the affective space.

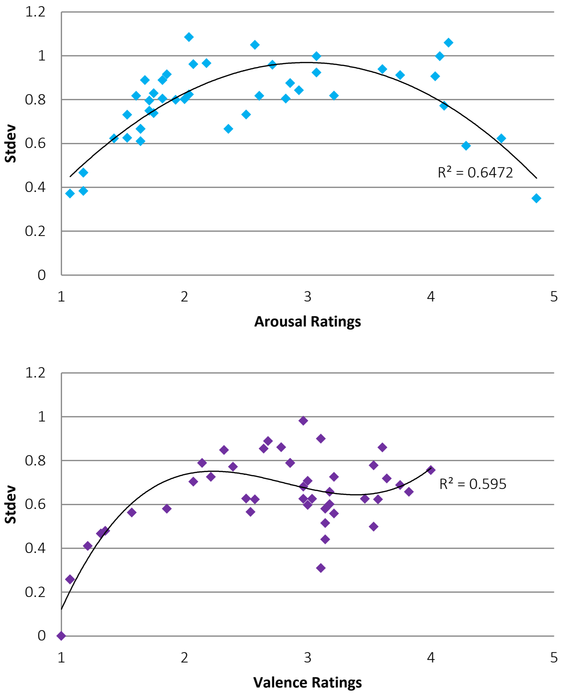


Figure 4: Standard deviations for both arousal and valence ratings for the 46 films clips and the associated best third-degree polynomial fitting curves. The coefficient of determination of the trend-lines is also indicated.

A t-test revealed that the Spearman's rank correlation coefficient (SRCC) between the rankings of the 46 film clips in the LIRIS-ACCEDE dataset and the ratings collected in this experiment exhibits a statistically highly significant correlation for both arousal ($SRCC = 0.751, t(44) = 7.635, p < 1 \times 10^{-8}$) and valence ($SRCC = 0.795, t(44) = 8.801, p < 1 \times 10^{-10}$). It indicates that the annotations gathered in an uncontrolled environment using crowdsourcing are highly correlated with the ratings gathered in a controlled environment. Fig. 3 shows that the excerpts selected for a specific axis are even more correlated with this axis than the other excerpts. Consequently, this new experiment in a controlled environment validates the annotations gathered using crowdsourcing that lead in [2] to the ranking of the LIRIS-ACCEDE dataset.

5.2 Discussion

The results from the experiment also exhibit a bias in the dataset to the extent that the distribution of the ratings for valence (see Fig. 2) shows that there are no film clips inducing high valence, which could be due to several factors.

First, it has been shown in previous works [1, 16] that positive evaluations were more subjective that negative ones. As a consequence, people agree more when they rate negative emotions than positive emotions. The plot of valence self-assessments corroborates the tendency that video clips that make people feel negative emotions elicit more consistent ratings than those that make people feel positive emotions (see Fig. 4). Since the ground truth is obtained by averaging subjects' ratings, the larger the standard deviation, the smoother the final value to a neutral value. In contrast to valence, Fig. 4 shows that the standard deviations of the ratings of the video clips eliciting extreme arousal (calm or excited) are lower than neutral ones.

Second, this bias may also be due to the fact that no movie from which the 9,800 excerpts included in LIRIS-ACCEDE are extracted induces high valence. However, 15% of the excerpts (1,477 film clips) in LIRIS-ACCEDE have been extracted from 25 comedy films. Major genres represented in the dataset are drama (28%), action/adventure films (16%), comedies (15%) and documentaries (14%) which are representative of current most popular movie genres. Contrarily to other affective video databases [9, 22], the excerpts in the dataset have been automatically segmented and thus have not been preselected in order to cover the whole affective space. But it seems highly unlikely that no excerpt or at least no scene in the selected movies induces high valence. Finally, another explanation of this bias is that it may be more challenging to induce very positive emotions in a short time than negative emotions. Indeed, the length of excerpts in the LIRIS-ACCEDE dataset varies from 8 to 12 seconds which may not be sufficient to elicit very positive emotions.

The rating experiment also reveals that the dataset suffers from another bias in the sense that there are less film clips with positive valence inducing high arousal, making the dataset asymmetrical. This bias can be found in other databases such as the EMDB dataset introduced by Carvalho *et al.* [5] that claimed that it is related to the existence of stronger response and attentional allocation to negative stimuli [7]. However, Lang *et al.* showed that the sexual stimuli included in the IAPS dataset elicited the most arousing and positive emotional reactions [12]. Because of ethical concerns, such sexual content is not included in the publicly available LIRIS-ACCEDE dataset, which might also partially explain the lack of highly arousing and positive content in the dataset.

6. CONCLUSIONS

This paper addressed the validation of large crowdsourced data and especially the LIRIS-ACCEDE affective video dataset. Following the work began in [2] in which the 9,800 film excerpts of the dataset have been ranked along the arousal and valence axes, the next step was to cross-validate the annotations gathered using crowdsourcing. Thus, we have proposed a different protocol consisting in collecting ratings for a subset of the dataset using the Self-Assessment Manikin (SAM) scales in a controlled setup. This subset consists of 40 excerpts that have been carefully selected based on their reliability to induce emotions during the crowdsourced experiment. The results have shown that the correlation between affective ratings and crowdsourced rankings is significantly high thus validating the overall dataset for future uses in research works.

Based on these results, we have been able to enrich the LIRIS-ACCEDE database by providing in addition to video rankings that were already available, video ratings thanks to a regression analysis that allows mapping all the 9,800 video clips included in the dataset into the 2D valence-arousal affective space [3]. In a near future, we plan to compare these ratings with continuous annotations made on longer video segments. We also intend to use this dataset to build a computational model taking into account temporal characteristics of movies and human emotions to automatically estimate the affective impact of videos.

7. REFERENCES

[1] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs. Bad is stronger than good. *Review of General Psychology*, 5(4):323–370, 2001.

[2] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret. A large video database for computational models of induced emotion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 13–18, 2013.

[3] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. From crowdsourced rankings to affective ratings. In *1st International Workshop on Multimedia Affective Computing (MAC)*, July 2014.

[4] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, Mar. 1994.

[5] S. Carvalho, J. Leite, S. Galdo-Álvarez, and O. Gonçalves. The emotional movie database (EMDB): a self-report and psychophysiological study. *Applied Psychophysiology and Biofeedback*, 37(4):279–294, 2012.

[6] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 491–500, 2009.

[7] B. L. Fredrickson. What good are positive emotions? *Review of General Psychology*, 2(3):300–319, 1998.

[8] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, Feb. 2005.

[9] 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*, 3(1):18–31, Jan. 2012.

[10] K. Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, Apr. 1970.

[11] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. *International affective picture system (IAPS): Technical manual and affective ratings.* The Center for Research in Psychophysiology, University of Florida, 1999.

[12] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm. Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, May 1993.

[13] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi. A supervised approach to movie emotion tracking. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2376–2379, May 2011.

[14] D. McDuff, R. E. Kaliouby, and R. W. Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012.

[15] S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, Aug. 2013.

[16] G. Peeters and J. Czapinski. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology*, 1(1):33–60, Jan. 1990.

[17] P. Philippot. Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition & Emotion*, 7(2):171–193, 1993.

[18] J. A. Redi, T. HoÃ§feld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel. Crowdsourcing-based multimedia subjective evaluations: A case study on image recognizability and aesthetic appeal. In *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, pages 29–34, 2013.

[19] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419, May 2011.

[20] L. Riek, M. O'Connor, and P. Robinson. Guess what? a game for affective annotation of video using crowd sourcing. In *Affective Computing and Intelligent Interaction*, volume 6974, pages 277–285, 2011.

[21] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003.

[22] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition & Emotion*, 24(7):1153–1172, Nov. 2010.

[23] J. A. Sloboda. Empirical studies of emotional response to music. In *Cognitive bases of musical communication.*, pages 33–46. American Psychological Association, 1992.

[24] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)*, pages 4–8, July 2010.

[25] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *IEEE Transactions on Multimedia*, 16(4):1075–1089, June 2014.