Is affective crowdsourcing reliable?

Isabelle Hupont
Aragon Institute of Technology
Email: ihupont@ita.es

Pierre Lebreton
T-Labs/Berlin University of Technology
Email: pierre.lebreton@telekom.de

Toni Mäki
VTT Technical Research Centre of Finland
Email: toni.maki@vtt.fi

Evangelos Skodras
University of Patras
Email: evskodras@upatras.gr

Matthias Hirth
University of Würzburg
Email: matthias.hirth@informatik.uni-wuerzburg.de

***Abstract*—Affective content annotations are typically acquired from subjective manual assessments by experts in supervised laboratory tests. While well manageable, such campaigns are expensive, time-consuming and results may not be generalizable to larger audiences. Crowdsourcing constitutes a promising approach for quickly collecting data with wide demographic scope and reasonable costs. Undeniably, affective crowdsourcing is particularly challenging in the sense that it attempts to collect subjective perceptions from humans with different cultures, languages, knowledge background, etc. In this study we analyze the validity of well-known user affective scales in a crowdsourcing context by comparing results with the ones obtained in laboratory tests. Experimental results demonstrate that pictorial scales possess promising features for affective crowdsourcing.**

I. Introduction

Affective Computing aims at developing autonomous systems able to detect and react to the users' emotions [1]. Such automatic affect detection algorithms (from voice, facial expressions, images, etc.) usually make use of supervised machine learning techniques which require labeled training data as the ground truth. The performance of these learning techniques depends on the quality of the training data and therefore on the quality of the labels. Additionally, in the User Experience (UX) field, more and more frequently content creators, web page designers or application developers ask for a more subjective, emotional state- and user engagement-related assessment from the users to complement other "traditional" objective metrics (such as number of application downloads, video views or tasks success rate).

In both fields, UX and Affective Computing, users affective information is usually obtained through emotional assessment right after the interaction. Several emotional questionnaires and scales are available in the literature such as AttrackDiff [2] or Differential Emotions Scale [3]. However, the users usually find it difficult and laborious to fill-in such tests, and given their verbal nature, linguistic issues may arise. Pictorial affective scales try to overcome these problems. Pictorial scales and their cartoon-like drawings enable people to visually express or report their emotions. One example is the well-known Self-Assessment Manikin (SAM) [4], a 9-point pictorial scale in two different emotional dimensions: pleasure (negative/positive) and arousal (passive/active). Pick-A-Mood (PAM) [5] is also a 9-point scale pictorial instrument for reporting and expressing not only emotions, but also moods.

Manual assessment by experts is the primary way of getting emotional information, but it can be an expensive and time-consuming process [6], and in many cases the generalizability of the subjective opinion of a small number of experts who often disagree [7] can be questioned. It has been demonstrated that emotional expertise does not necessarily correlate with emotional experience [8], suggesting that wider non-expert annotations are needed to obtain a realistic affective feedback. One possible way to overcome the aforementioned issues is to outsource the task to a large group of non-expert (international if so wished) individuals. Crowdsourcing, a practice of dividing labour between large number of (typically online) workers, is a promising method for such outsourcing [9]. One of the recent applications of crowdsourcing has been to label training data for a wide range of supervised learning application domains. Practical experiences with crowdsourcing have demonstrated that it can offer a fast, cheap and effective way to collect labels [10]. However, most of these experiments have been carried out for annotating data of objective nature.

The annotation and assessment of emotions through crowdsourcing has not been yet fully explored [10], even though it could bring interesting information not just in terms of the number of annotations but also with respect to the cultural differences in the perception/expression of affect. There are few projects that focus on the intersection of affect and crowdsourcing. The framework by McDuff et al. [11] allows to automatically collect and analyze facial responses (smiles) to media contents massively over the internet. Soleymani et al. [12] propose a platform for the affective annotation of videos in terms of one of 10 emotional labels. Finally, the works by Tarasov et al. [8] and Morton et al. [13] explore the audio channel to emotionally annotate human speech and music, respectively. However none of those studies provide insights about the cultural differences found in the annotations or the reliability of the crowsourced emotional data.

Regarding the issue of reliability, current research in crowdsourcing shows that the number of untrustworthy users is usually not large [8]. There is evidence of a number of different techniques ("honeypots") used to guard against malicious or lazy test participants: some research requires users to show some degree of accuracy on a small test subset [14] while other works include explicitly verifiable questions to reduce invalid responses [15]. With respect to the reliability in the annotations themselves as ground truth data, several works have demonstrated high inter-agreement among crowdsourcing annotators, similar to the one obtained in a laboratory environment [16].

While this is true for objective annotation tasks, is it also possible to rely on affective crowdsourcing? The problem of massive emotional assessment is particularly challenging in that it may change from persons to persons due to different social and educational background, and even different cultures or languages. The main scope of this work is to study how affective assessment can be evaluated in a crowdsoucing test.

The remainder of the paper is organized as follows. We first present and discuss our starting point, the work by Dan-Glauser and Scherer [17], that recruited participants in a laboratory environment to emotionally annotate a set of highly affective pictures (Section II). Then we explain how we brought the same experiment to an international crowdsourcing environment (Section III), and we analyze the differences obtained in terms of annotations and the origin of the problems arisen (Section IV). Finally, we conclude by presenting interesting insights for affective crowdsourcing (Section V).

II. Background and Objectives

A. The Geneva Affective PicturE Database

The Geneva Affective PicturE Database (GAPED) was constructed with the aim to provide a well-characterized and salient visual stimuli that can be reliably used as affect inducing material [17]. It consists of 730 pictures divided into six categories. The negative contents (4 categories) chosen are spiders, snakes, scenes concerning human rights violation and animal mistreatment. The positive category involves human and animal babies as well as nature sceneries, whereas the neutral category pictures mainly depict inanimate objects.

Each picture in the database is annotated in terms of emotional valence and arousal. The valence measures how a human feels, from positive to negative, while the arousal measures whether humans are more or less likely to take some action under the emotional state, from active to passive. These annotations were obtained after several rating sessions in a laboratory environment with a total of 60 participants. During the sessions, each picture was presented in full screen to the participants for 4 seconds. After the presentation, continuous colored rating scales (from 0 to 100) were shown on a new screen (along with a small-sized exemplar of the picture as a reminder) and the participants could begin rating. The valence scale went from *negative* to *positive* and arousal scale from *calm* to *excited.*

B. Moving affect assessment to crowdsourcing environments

A few limitations about the GAPED construction have to be discussed. Firstly, the participants were recruited from a second-year psychology class, which implies that the vast majority of them had a clear understanding of the emotional models used and the concepts of arousal and valence. Secondly, all of them were perfectly fluent in the language in which the study was conducted and possible questions and ambiguities could be addressed in person before starting the test. Both assumptions (affective models familiarity and language fluency) are not realistic in crowdsourcing environments.

Unconstrained environmental and multi-cultural conditions render gathering affective information challenging. As opposed to the lab experiments, crowdsourcing environments are almost impossible to properly control and the influence of exogenous distracting factors cannot be quantified or measured. The material's emotional power may be suppressed by the uncontrollable presence of distracting or emotion inducing factors. Crowdsourcing campaigns can be launched globally making them available to people coming from different cultures and speaking different languages. The perception of emotional contents may be affected by cultural and educational factors. Moreover, highly subjective tasks of emotion assessment do not have exact wrong answers. Thus, distinguishing untrustworthy or lazy participants is challenging, requiring a meticulous selection of different quality control techniques (e.g. "honeypots").

This work aims to investigate how crowdsourcing can be used to extend the GAPED annotation, enabling more diversity in test participants. To this end, we have defined and launched different crowdsourcing campaigns as explained in the following sections.

III. Affective crowdsourcing: experimental setups and campaigns description

The crowdsourcing campaigns were carried out using the Microworkers[1] crowdsourcing platform. Two different experimental setups were developed and used in four different campaigns. A first setup and campaign were used to guide the second revision of the setup and latter three campaigns.

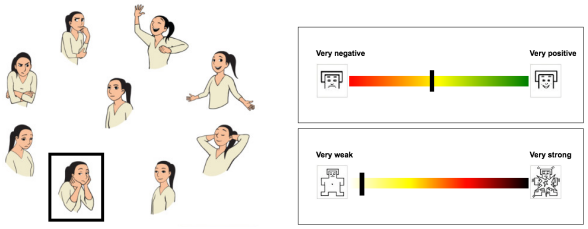
A. First setup: The original experiment adapted to crowdsourcing

The initial setup was an attempt to repeat the original experimental conditions in a simple and straightforward design with minimal changes. A necessary addition to the design, because of our research question, was the inclusion of the PAM pictorial affective scale [5]. We selected a subset of 180 images from the original database, consisting of the 30 most "emotion inducing" pictures from each category (images with higher valence/arousal values per category).

At first, the participants were given instructions regarding the aim of the study and the tasks to accomplish, including a disclaimer about possibly shocking content. Subsequently, the rating scales (c.f. Figure 1) were presented in detail and anonymous personal information regarding the gender, ethnicity and age were collected. In the actual test each user was asked to rate 7 pictures using both scales. Six of the displayed images were randomly selected, by drawing a single image from each category. The remaining image was one of these six pictures (randomly chosen), that was repeated as a consistency check, in such a way that the picture never appears two times consecutively. Considering the number of test participants (30), this resulted in one annotation per image.

Each picture was displayed in full size for 4 seconds. Then the users were presented a question about the content of the picture ("honeypot"). Subsequently, participants rated their mood using the PAM scale, answering the question *How do you feel after looking at that image?*. They then rated their experienced levels of valence and arousal using two 0 to 100 scales. Valence was rated by choosing a value from scale of

[1] http://www.microworkers.com



(a) Pick-a-Mood (PAM) scale. (b) Valence and Arousal scales.

Figure 1: Rating scales during the first campaign.

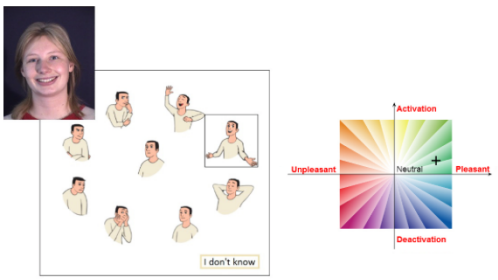


Figure 2: Rating scales example from the second campaign.

Very negative to *Very positive* as response to the prompt *Click on the color bar according to how you feel about the picture*. The arousal was rated with a scale from *Very weak* to *Very strong* as a response to the prompt *How strong is your feeling?*. There was a Self-Assessment Manikin [4] at each end of the bar to assist users (Figure 1b). A downsized version of the picture being rated was visible on both rating pages. There was no time limit to complete the ratings. After the ratings, the test continued with the next picture.

At the end of the test, the participants were given the option to leave feedback and also throughout the test there was a possibility to ask for help using a dedicated discussion forum. After completing the post-test questionnaire, they received the token to be submitted using Microworkers site in order to receive the payment.

B. Second setup: Making the tasks easier to understand

After analyzing the results of the first campaign (c.f. Section IV), and taking into account the feedback received from the users, we made several improvements to the setup. Valence and arousal scales were replaced by a 2-dimensional space, with valence on X-axis (from *Unpleasant* to *Pleasant*) and arousal on Y-axis (from *Deactivation* to *Activation*). We also included some examples to demonstrate the connection between expressions and ratings (c.f. Figure 2) and, after that, an interactive training phase was added to acquaint the participants with the scales and concepts. Finally, an extra *I dont know* option was added to the PAM scale as some users found the options insufficient (e.g. some of them reported to miss a "disgust" mood picture).

Additionally, we also re-designed a set of factors to focus the user attention in the images emotional annotation task itself. Firstly, the question about the content appeared only once, randomly after one of the 6 images, since it was assumed that the surprise element would make answering the question difficult for a user not committed to the task. Secondly, we modified the consistency check so that the users were explicitly asked to reproduce similar scores while rating the repeated image, presuming that a big change in annotations could be attributed to a user not concentrated in the task.

From a more technical point of view, in this second setup the number of images per category was also decreased from 30 to 5 (i.e. a total of 5x6=30 images from the original database were used). This allowed to collect at least 6 annotations per image and therefore to apply certain statistical methods to analyze the results.

C. Towards affective crowdsourcing: Campaigns roadmap

The 1st campaign's main objective was to validate the original test setup. The users were invited globally to get quick feedback. It took only 25 minutes to receive the required 30 completed assessments. The 2nd campaign was launched after implementing the improvements (c.f. Section III-B). One conclusion from the first campaign was that the comprehension of the language used may have played a role. Consequently, the second campaign was made available only in the US to recruit a larger amount of native English speakers. This time it took two and half weeks to receive 30 completed assessments.

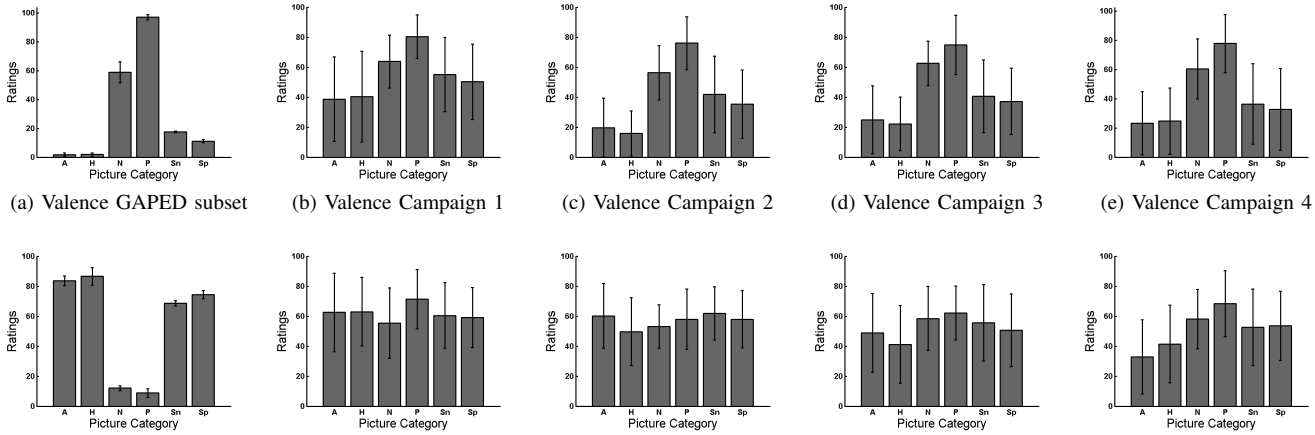
While the results from the 2nd campaign were promising, they left the question of language and culture effect open. To this end, the 3rd and 4th campaigns were targeted towards two expectedly different cultural and lingual areas, a set of Asian and a set of European countries, respectively. It took less than 1 day to obtain the results from both campaigns. According the the guidance by Microworkers the participants were paid reward varying from 30 to 40 USD cents depending on the country of origin.

IV. Campaigns Results

A. General results

Figure 3 shows the mean valence (first row) and mean arousal (second row) per picture category for the in-lab experiment and the crowdsourcing campaigns. As it can be seen, the laboratory results have narrower confidence intervals than the crowdsourcing results. Several reasons can explain the differences found. Firstly, GAPED results include ratings from *all* the experts for *each* picture. In the crowdsourcing tests the subset of images chosen represented the most "emotional" ones and there were fewer ratings per picture. Secondly, given the high concept awareness and expertise of the laboratory test subjects, better agreement in annotations was highly expected from the laboratory results.

Concerning the valence, the comparison between the in-lab and the 1st campaign results (Figures 3a and 3b) reveals the latter's relatively low performance, which can be attributed to the low number of annotations per image and the first setups weaknesses. The results from the 2nd, 3rd and 4th campaign (Figures 3c- 3e) resemble more closely the in-lab results form of Figure 3a. This would imply that the notion of valence was better conveyed with the second setup. The large confidence intervals for all crowdsourcing campaigns were expected, given



(a) Valence GAPED subset (b) Valence Campaign 1 (c) Valence Campaign 2 (d) Valence Campaign 3 (e) Valence Campaign 4

(f) Arousal GAPED subset (g) Arousal Campaign 1 (h) Arousal Campaign 2 (i) Arousal Campaign 3 (j) Arousal Campaign 4

Figure 3: Mean valence and arousal for the GAPED subset used, per campaign and image category - (A)nimal mistreatment, (H)uman, (N)eutral, (P)ositive, (Sn)akes and (Sp)iders.

the highly subjective nature of the task. Regarding arousal, the results are ambiguous, manifested as almost random mean values combined with low agreement between the annotators. This could be attributed to the fact that arousal is more difficult to conceive, define and express intuitively than valence. The difficulty to draw concrete results using valence/arousal scales, supports the idea of using other emotional scales such as PAM.

Figure 4 presents the confusion matrices with respect to PAM and emotion categories for the different crowdsourcing campaigns. We can observe agreement in annotations inside continents, which demonstrates the advantage of using pictorial scales for crowdsourcing tasks. Such scales, using a basic, limited set of pictures of emotions common for all humans regardless of their culture and origin, seem to alleviate the difficulties in expressing affective perceptions.

B. First setup results in-depth

Regarding the 1st campaign, we started by studying the quality of the subjective data. The test participants' commitment was evaluated by checking the questions about image content. From the 29 participants, 21 answered correctly to all questions and 8 gave one false response. The results are satisfactory, but a possible issue was identified: the surprise effect may have disappeared after the first question and the test participants may have started to focus on image's content, lowering the quality of the emotional ratings.

A second data quality check was the evaluation of repeated images. A possible issue was identified also with this technique: it is possible that the participant's emotion changes between repetitions, e.g. after having seen the other pictures. Therefore it was concluded that no users should be rejected based on this implementation of the repetition and a revised design was implemented for the second setup (c.f. Section III-B).

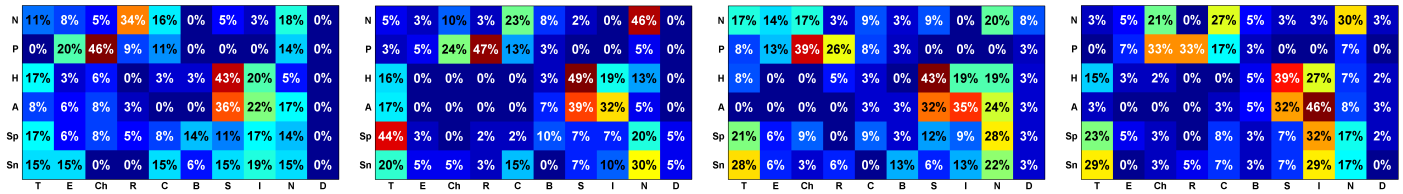
A comparison of the original annotations from [17] and the crowdsourcing campaign (c.f. Figure 3) plus the observed noise in the data revealed the low performance of the first crowdsourcing implementation regarding valence and arousal. The low performance was potentially the result of three root causes: the participants were not familiar with the vocabulary used or with the concepts used, and the scales may have not been clearly understood.

C. Second setup results in-depth: Expanding to cross-cultural and multilingual settings

As in the previous section, the content question was checked to analyze the quality of the subjective data for the 2nd, 3rd and 4th campaign (second setup). From the 30 test participants, only 3 did not answer it correctly. However, considering participant's eligibility included also the evaluation of the repeated picture, which indicated consistent ratings by all users. In the end no users were rejected.

The relationship between the valence scores from the different experiments involving the American, Asian and European test participants was analyzed. The observed Pearson correlations between valence ratings given by the different groups were: 0.86 between US and Asian, 0.86 between US and European and 0.80 between Asian and European participants. This implies that the subjective ratings were consistent between experiments and the comprehension of English did not play a big role. A one-way ANOVA was applied to the data to investigate further the effect that groups (different campaigns) and countries of origin have on the valence scores. The ANOVA shows that there is no significant effect of the groups on the ratings at a confidence interval of 95% ($F=1.2$, $p=0.34$). A non-parametric Friedman test was executed to consider the effect of experiment per category of images and no significant effects were identified. The effect of the country of origin to valence was also analyzed with ANOVA and found to be moderately significant ($F=1.9$, $p=0.05$).

A similar analysis was performed regarding the use of the arousal scale and the correlation between the groups' votes was found low: 0.19 between US and Asian, 0.36 between Asian and European, and -0.11 between US and European votes. The correlation between Asian and European ratings seemed to be



(a) Campaign 1 (b) Campaign 2 (c) Campaign 3 (d) Campaign 4

Figure 4: PAM confusion matrices for the different campaigns, showing the picture categories - (N)eutral, (P)ositive, (H)uman concerns, (A)nimal mistreatments, (Sp)iders, (Sn)akes- against different moods they evoke - (T)ense, (E)xcited, (Ch)eerful, (R)elaxed, (C)alm, (B)ored, (S)ad, (I)rritated, (N)eutral, (D)o not know.

higher, however the values are too low to be conclusive. An ANOVA analysis was performed also on the arousal scores. The results didn't show a significant effect of the experiment on neither the arousal scores in general (F=0.9, p=0.43) nor the arousal scores per category. Also the effect of the country of origin on the arousal score was not found significant (F=1.3, p=0.25).

On the other hand, the PAM scale seems to highlight better the cultural differences among countries. As an example, Figure 5 illustrates the differences of votes acquired from different nationalities assessing pictures about spiders using the PAM scale. Interestingly, in India, Pakistan and Bangladesh participants felt calm, neutral or bored, while in many other countries (mainly in US and Europe) irritation or tension were mostly reported. Finally, for all the different campaigns, animal and human mistreatment induced strongly negative feelings, however, the 'sad' feeling was more dominant for human concerns while 'irritated' was more dominant for animal mistreatment. This observation may imply that users feel more actively involved when confronted with animal mistreatment, but results are limited to the number of test participants and it is hard to strongly tell if differences come from individuals or cultural issues.

D. Crowdsourcing vs. lab emotional scales comparison

Combined results from the 2nd, 3rd and 4th crowdsourcing campaigns were analyzed. First, the mean valence and arousal values and the dominating mood for each picture were calculated. Then the Valence/Arousal (VA) results from laboratory and crowdsourcing campaigns were converted to PAM ratings. The transformation exploited the design of the PAM scale,

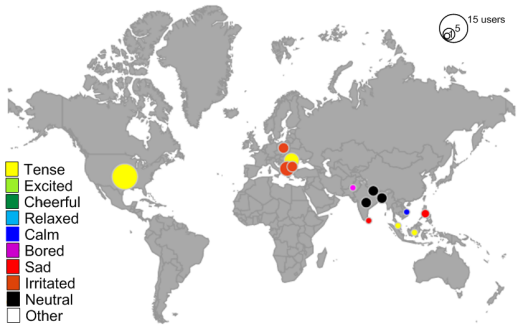


Figure 5: Cultural differences in PAM ratings for "spiders".

Table I: Agreements (in percentages) of laboratory and crowdsourcing originated annotations.

	Relaxed comparison (30 pictures)	Strict comparison (30 pictures)
CS_{PAM} vs. $GAPED$	83.3%	23.3%
CS_{VA} vs. $GAPED$	53.3%	11.1%
Significance of difference (P -value)	0.03	0.29

where the specific moods are located according to the valence and arousal values (VA points) interpreted as coordinates (valence as X-coordinate and arousal as Y-coordinate), as can be observed in Figure 1a. In this transformation VA points in the circle centered in origin are mapped to neutral mood of the PAM scale. The rest of the VA space is divided into 8 sectors, each represented by a single mood of PAM scale. Accordingly, the VA points located in a sector are mapped to the mood representing that sector.

Next, the annotations from the crowdsourcing campaigns and laboratory tests were compared for each picture present in both campaigns. Two strategies were applied in the comparison process: *strict comparison*, where the mood annotations had to match exactly; and *relaxed comparison*, where the compared moods had to either match exactly or be adjacent to each other. Table I summarizes the observed agreements between the annotations from crowdsourcing and laboratory studies. The annotations acquired with PAM scale are better aligned with the laboratory results (CS_{pam} vs $GAPED$) than the annotations acquired with Valence/Arousal (CS_{VA} vs $GAPED$). The difference between the scale's agreements was analyzed with a Chi-squared test and found statistically significant in the case of relaxed comparison.

To further investigate the differences of reliability between the VA and PAM ratings, we compared the confidence intervals of the scores from each methodology per group of pictures. To average the PAM ratings and to compare the confidence intervals, the mood values were converted to VA points according the method described in [5] (it provides a correspondence table between moods and average values of valence and arousal including standard deviations). Let M_{v_m} be the average valence value for the mood m, let V_{v_m} be the variance of corresponding to the mood m, and let N_m the number of subjective scores which have been used to determine the values of M_{v_m} and V_{v_m}. It is possible to determine the global mean valence score, GM_v, from the different moods Ω by averaging the average value of valence corresponding to each mood and weighting them by the number of score used

Table II: Comparison of confidence interval size between PAM and VA ratings.

	Campaign 1				Campaign 2-4			
	PAM		V/A		PAM		V/A	
Images	Mean	CI	Mean	CI	Mean	CI	Mean	CI
Animal	0.57	0.11	0.45	0.28	0.53	0.11	0.38	0.29
Human	0.57	0.11	0.46	0.27	0.56	0.11	0.43	0.29
Neutral	0.53	0.11	0.54	0.24	0.54	0.10	0.42	0.25
Positive	0.67	0.11	0.63	0.27	0.56	0.12	0.42	0.30
Snake	0.65	0.13	0.58	0.25	0.59	0.10	0.44	0.29
Spider	0.64	0.12	0.61	0.29	0.60	0.10	0.47	0.34

to get the average value: N_m. The global variance can be also determined as follows:

$$GV_v = \frac{\sum_{m\in\Omega} V_m \times N_m + (M_{v_m} - GM_v)^2 \times N_m}{\sum_{m\in\Omega} N_m} \quad (1)$$

The values of N_m were determined based on the annex section in paper [5] using the number of provided labels per mood. Based on these formulas, the global means and confidence intervals (in the VA space) per group of picture could be determined for the PAM ratings. This allowed us to compare PAM ratings to the ratings collected using valence and arousal scales. Table II shows the global means and confidence intervals for valence ratings for each category of pictures. The PAM ratings have smaller (less than half the size) confidence intervals than the VA ratings, which indicates a higher consistency between scores when using the PAM scale.

V. Conclusions

In this paper the reliability of collecting emotional data using crowdsourcing was evaluated. Two different setups were employed, the first of which was a straightforward implementation of the original in-lab design, while the second setup comprised several adaptations in order to meet the specificities of crowdsourcing. The main difficulties which were encountered stemmed from the subjective nature of emotion assessment, making it difficult to check the commitment of the users and their level of understanding of the task. The wide demographic scope, involving a variety of people with diverse attributes regarding their social and educational background, and even different cultures and languages, rendered the problem more challenging.

The comparison with in-lab experiments demonstrated promising results regarding valence, while for arousal the results were more ambiguous, due to the increased difficulty in conceiving and defining it. Using the pictorial PAM scale, more consistent results among the annotators were reported. The comparison of different emotion evaluation scales supports the hypothesis of the superiority of pictorial scales compared to other affective scales. Possibly the universality of emotions depicted on pictorial scales make them better address the cultural and language differences. Concluding, although the issues of annotating and assessing emotions through crowdsourcing still remain open, the current work provides a first proof regarding its feasibility, providing also interesting insights. The significance of the current results is also reflected in the fact that crowdsourcing is a vast, growing field that offers a prospective opportunity for large-scale data annotation, given its speed and low cost.

Acknowledgments

This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET. T. Mäki's and I. Hupont's work was carried out in the context of the CELTIC Plus QuEEN project, partially funded by Tekes, the Finnish Funding Agency for Technology and Innovation, and the Spanish Ministry of Science and Innovation. The authors would like to express their gratitude to Microworkers for funding the embursement of the workers for their contributions.

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] M. Hassenzahl, M. Burmester, and F. Koller, "Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität," in *Mensch & Computer 2003*. Springer, 2003, pp. 187–196.

[3] C. E. Izard, *The psychology of emotions*. Springer, 1991.

[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*, vol. 25, no. 1, 1994.

[5] P. Desmet, M. Vastenburg, D. Van Bel, and N. Romero, "Pick-a-mood: Development and application of a pictorial mood-reporting instrument," in *Proceedings of the International Design and Emotion Conference*, London, UK, Sep. 2012.

[6] P.-Y. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: A study of annotation selection criteria," in *Proceedings of the Workshop on Active Learning for Natural Language Processing*, Boulder,USA, Jun. 2009.

[7] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, 2010.

[8] A. Tarasov, S. J. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *Proceedings of the Workshp on EmotionML*, Paris, France, Oct. 2010.

[9] R. Morris, "Crowdsourcing workshop: The emergence of affective crowdsourcing," in *Proceedings of the Workshop on Crowdsourcing and Human Computation*, Vancouver,Canada, May 2011.

[10] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Waikiki, USA, Oct. 2008.

[11] D. McDuff, R. Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," *Affective Computing, IEEE Transactions on*, vol. 3, no. 4, 2012.

[12] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proceedings of the Workshop on Crowdsourcing for Search Evaluation*, Geneva, Switzerland, Jul. 2010.

[13] B. G. Morton, J. A. Speck, E. M. Schmidt, and Y. E. Kim, "Improving music emotion labeling using human computation," in *Proceedings of the Workshop on Human Computation*, Washington,USA, Jul. 2010.

[14] V. Ambati, S. Vogel, and J. G. Carbonell, "Active learning and crowdsourcing for machine translation." in *Proceedings of the Conference on Language Resources and Evaluation*, Valletta, Malta, May 2010.

[15] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements," in *Proceedings of the Conference on Human Factors in Computing Systems*, Florence, Italy, Apr. 2008.

[16] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the Conference on Multimedia Information Retrieval*, Philadelphia,USA, Mar. 2010.

[17] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," *Behavior research methods*, vol. 43, no. 2, pp. 468–477, 2011.