Crowdsourcing for Affective-Interaction in Computer Games

Gonçalo Tavares, André Mourão, João Magalhães
Dep. Computer Science, CITI/FCT
Universidade Nova de Lisboa
Portugal
g.tavares@campus.fct.unl.pt, a.mourao@campus.fct.unl.pt, jm.magalhaes@fct.unl.pt

ABSTRACT

Affective-interaction in computer games is a novel area with several new challenges, such as detecting players' facial expressions (e.g., happy, sad, surprise) in a robust manner. In this paper we describe a crowdsourcing effort for creating the ground-truth of a large-scale dataset of images capturing users playing a computer game. The computer game is designed to elicit a particular facial expressions and the game will score the player according to the detected expression. For designing the crowdsourcing task, some of the examined variables include: reward, tagging limits, golden questions, workers' location. In the end, we designed a large tagging job to maximize workers agreement. Each image with a facial expressions is tagged with one of the following expressions labels: *happy*, *anger*, *disgust*, *contempt*, *sad*, *fear*, *surprise*, and *neutral*. The dataset included over 40,000 images, the workers' judgments, the game's detected facial expression and what facial expression the player should be performing.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Crowdsourcing, affective-interaction, facial expressions.

1. INTRODUCTION

The scientific interest in facial expressions was first systematized by Darwin in its seminal work "The expression of emotions in man and animals" [5]. Since then, others have researched the area and more recently studied the link between emotion and facial expressions. Ekman et al. [7] identified six primal emotions that can be mapped into facial expressions. Facial expressions are a key part of human communication - they complement natural language and gestures. As Tian, Kanade and Cohn [14] put it, facial expressions are "the facial changes in response to a person's internal emotional states, intentions, or social communications". Interest in facial expressions has grown steadily in computer vision. They are crucial for building new systems with affective interaction capabilities. Computer games are a primary example, of where game actions can be adjusted to the payer's facial expression. To design such applications, one needs data resources to research and tune the required technology. In this paper we describe a crowdsourcing job task for gathering such resources. Image data for affective-interaction was collected through a gamification process. A game was implemented [4] to capture player's faces while interacting with the game. A large set of unlabeled interaction images were collected. Next, a crowdsource process was used to annotate each image with a facial expression. Since we collected a large number of images it was not possible to have them annotated by an expert or professional. Thus, we resourced it to a crowdsourcing service. The main contribution of this article is a dataset obtained in the following two steps:

- **Gamification** was used to collect image data of real players engaged in a computer game. Most importantly, players were aware that their facial expression was in full control of the game-play.
- **Crowdsourcing** was used to tag the facial expression of all images. The large-scale crowdsource tagging effort covered a set with over 40,000 images and 6 emotions. The judgements were collected for the facial expression labels and intensity.

The design of the crowdsource job was carefully planned: we obtained several judgments per image, each facial expression was linked to an intensity, and different worker selection criteria and batch jobs were inplace to reduce bias. This effort is also particularly relevant because, the annotations of a facial expression image will not be a binary label, but a distribution across the different expression labels. This is extremely useful for creating better affective-interaction models.

The remainder of this paper is organized as follows: Section 2 describes the related work, section 3 details the gamification process (the acquisition of the facial-expressions interaction images), and section 4 details the crowdsource process. Finally, section 5 discusses the judgment results and the corresponding conclusions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CrowdMM'13, October 22, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2396-3/13/10 ...$15.00.
http://dx.doi.org/10.1145/2506364.2506369.

2. RELATED WORK

2.1 Gamification

Although the benefits of tagging multimedia are immense, exaustively capturing and tagging content can be cumbersome. Ames et al. [2] analysed motivations and incentives for tagging photos through the usage of games (e.g. ESP game [16] and ZoneTag [1] are some of the examples). In [2] it is hypothesized that multiple motivations are a determining factor in the users decision to annotate - especially social incentives. For the dataset described in this paper, a game was designed to capture images and crowdsourcing was used to label the images with the correspondig facial expression.

2.2 Crowdsourcing

The progress and expansion of the Internet allowed every person to submit sets of micro-tasks and to undertake those micro-task. Crowdsourcing sites made this process automatic in many ways (see [19] for a survey of crowdsourcing). These sites support two types of users: *requesters*, who submit jobs (sets of micro-tasks), and *workers*, those who perform the micro-tasks. Financial rewards are central to the success of crowdsourcing jobs: they are the motivation for workers to complete their micro-tasks with the best possible quality. To achieve the best quality judgments (the result of a micro-task), one must be aware that different workers produce different quality results. The main reason for a worker to undertake a job is the financial reward. However, there are other factors that contribute for the reliability of the results. One common approach is to design the crowdsourcing job as a game, allowing the worker to engage in a micro-task as an entertainment activity. Other approaches appeal to the worker's will to contribute to a greater cause, such as annotating computer vision medical data [8]. Thus, incentives are not only monetary, but also entertainment-wise or for a noble-cause. Besides the "incentive" aspects, to improve results' quality, the job designers must deploy several quality control measures. At the job-design phase, the requester introduces some gold questions into the micro-task workflow to characterize workers' trust and eventually single-out spammers [12]. Later, at the end of all micro-tasks, the aggregated results generate an agreement that allows assessing the quality and reliability of individual workers [17]. Crowdflower[1] is the crowdsourcing site used in our work, this site follows a workers/tasks policy different from other competitors, such as Amazon's Mturk [2]. Instead of having their own workers, Crowdflower also relies on a network of other crowdsourcing partners - Mturk is one of such partners.

2.3 Crowdsource and training data

Several research areas have their own definition of relevance giving more emphasis to their specific objectives: Information Retrieval aims at finding documents that best answers a particular user query, Computer Vision aims at detecting image objects or contexts. All related areas rely on datasets of labeled examples, whose relevance was judged by a human. Unfortunately, it is hard to define relevance formally and universaly: the notion of a relevant item is diffuse because the same item can have different interpretations to different humans. These discrepancies are more noticeable in large multimedia collections for two reasons: (1) multimedia information is not as concrete as textual information, thus more open to different interpretations and relevance judgments (types of relevance); (2) assessing the relevance of documents is an expensive task requiring human-effort for long periods of time, thus, collections with a large number of documents are only partially annotated: most collections are incomplete and inconsistent. To overcome the burden of annotating data for computer vision, researchers look at alternative solutions that reduce the human effort through either automated methods [10] or crowdsource solutions to generate labeled data [13].

2.3.1 Types of judgements

According to the information domain, different definitions of relevance are more adequate than others. Three types of relevance judgements are easily identified in the literature:

- **Binary relevance:** under this model a document is either relevant or not. It makes the simple assumption that *all* relevant items contain the same amount of information value.
- **Multi-level relevance:** one knows that documents contain information with different importance for the same query, thus, a discrete model of relevance (e.g., relevant, highly-relevant, not-relevant) enables systems to rank documents by their relative importance.
- **Ranked relevance:** when documents are ordered according to a particular notion of similarity.

The binary relevance model is the most common practice in HCI and AI systems. These systems are tuned with a set of judgments that reflect the majority of experts' judgments. The multi-level relevance provides the annotator with more expressive power than with binary relevance - e.g. workers feel more confortable with three or four levels of relevance-intensity instead of only true/false. The relevance judgments of the ranked relevance model are actually a rank of documents that exemplify the human perception of a particular type of similarity, e.g., texture, colour. In practice, for the task at hand, only the binary or the multi-level judgments are viable.

2.3.2 Judgments' quality

The judgments quality of crowdsourcing jobs has been the matter of much research, [11]. Traditionally, expert annotations are obtained through processes that eliminate problems of inconsistency and bias. Volkmer et al. [15] followed the following rules to improve judgments' quality: (1) assessors annotated a sub-set of the documents with a sub-set of the labels (this avoids the bias caused by having the same person annotating all data with the same concept); (2) all documents must receive a relevance judgment from all annotators (this eliminates the problem of incomplete relevance judgments but increases inconsistency); and (3) documents and labels were assigned to annotators so that some documents received more than one relevance judgment for the same label; this eliminates the inconsistency problem if a voting scheme is used to decide between relevant and non-relevant.

Vokmer's et al. [15] annotation study was a quite formal and expensive processes. Nowak et al. [11] compared the

[1]http://www.crowdflower.com/
[2]http://www.mturk.com/

judgments quality of expert to that of individual workers. They confirmed through several statistical measures, that when considering the aggregated results of the non-experts (crowdsource workers) the judgments were comparable to the ones created by experts.

3. AFFECTIVE-INTERACTION DATA

This section will describe how the affective interaction data was captured. It's a two-players game where the objective is to perform a set of facial expressions. Players play simultaneously and facial expressions are competitively scored. The player that performs an expression closer to the one asked, wins a (timed) round – the player who wins more rounds wins the game. The game is described in detail in [4].

Facial expressions are represented by a label and a related image. Players are instructed to guide themselves by the label to avoid ambiguity. Figure 1 displays the main game interface. The players are trying to perform the *disgust* expression. The colored bars represent the scores (top: score of the last image, bottom: best score of the round); the numbers at the center represent the global scores; the half circle is the round timer; the image and text at center represent the reaction image and label; the faces at left and right are from the players and the label represents the last expression recognized – see [3] for details of the facial expressions analysis algorithm.

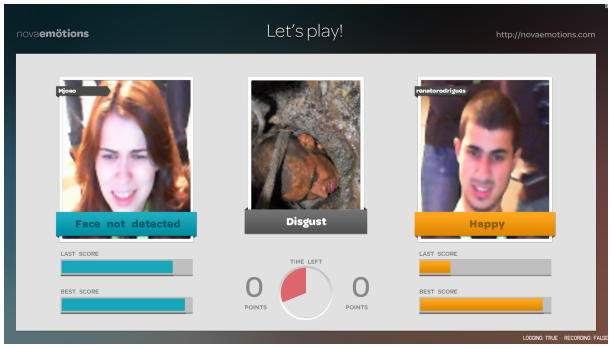


Figure 1: The game interface.

3.1 The subjectiveness of a facial expression

Humans are able to recognize different facial expressions and infer what emotion that expression conveys. Ekman [6] defined a total of six basic universal emotion expressions: *Happiness*, *Sadness*, *Surprise*, *Fear*, *Anger* and *Disgust*. *Neutral*, a state of no visible expression, and *Contempt*, a mixture of *Anger* and *Disgust* are also part of Ekman's suit of facial expressions of emotions. These are the expressions we have chosen in our work. But changes in facial expression can be more subtle like moving the outer section of the brows or depressing the corners of the lips. The expressions described above can be defined as a set of Action Units (AUs). An AU is an action performed by one or more muscles of the face humans are able to distinguish. A full description is available at Tian et al. [14]. Even using AUs, some expressions are similar: *Fear* is composed by the same AUs as *surprise* plus other 3 AUs [7].

In a real world situation, the differences between facial expressions are even more subtle. A person can be performing ambiguous expressions (e.g. *surprise* and laughter) or performing unrelated actions (e.g. talking) at the same time. This poses a problem for classification. In our case, images come from a game where people can feel awkard or are simply adjusting themselves to the game control. Thus, our images are from a real setting where people are interacting with their faces, and consequently several images are ambiguous or laught. To account for these situations we included labels the labels *ambiguous* and *not a face*.

In Ekman's work, expressions are clearly defined as AUs, but we assumed that workers are not familiar with the AU based composition of the facial expressions. We rely on a large sample of judgments to provide us with a reliable set labeled data.

3.2 Data acquisition setting

We captured over 40,000 facial expressions during the game trials. These face images were captured in a novel and realistic setting: humans competing in a game where players' facial expression have an impact on the game. Some example faces with labels are visible on Figure 2. These images offer a novel view of facial expression datasets: players were competing using their own facial expressions as an interaction mechanism, instead of performing well defined prototype expression.

This dataset is also unique in the following senses: user faces are not in fixed positions (about 50% of the face images are not front facing and are at different heights). Existing facial expressions datasets like CK+ [9] or the BU-4DFE [18] datasets were captured in controlled environments and, in CK+, by people trained to perform a prototype expression.

Our approach was different: the volunteers were asked to perform an expression in a social gaming environment with varying lightning, background and position. Thus, a pure affective-interaction setting where the computer is controlled by the players' facial expression. Each captured image contains the information regarding the expected expression and the expression detected by the game algorithm.

4. CROWDSOURCING TASK DESIGN

Crowdsource task desingns go beyond the workers interface. Therefore, in this section we identify the factors that contribute directly or indirectly to achieve reliable and relevant results. These factors can be divided into two distinct groups, worker qualification and job attributes.

4.1 Worker qualification

To be accepted in a job, a worker must pass in some qualification criteria. The skills between workers are different, therefore we need to find the group of workers that best suits the micro-task. The qualification process of a worker is based on the following attributes:

Country. Each country has its own culture, customs and so on. Our criteria must ensure that the worker is capable of performing the micro-tasks based on his life experience. To ensure this, it's possible to include or exclude countries from the list of allowed countries. Thus, we favoured English speaking countries.

Judgment limits. The maximum number of judgments a worker is allowed to make can be limited: small maximum judgement values can increase significantly the duration of a

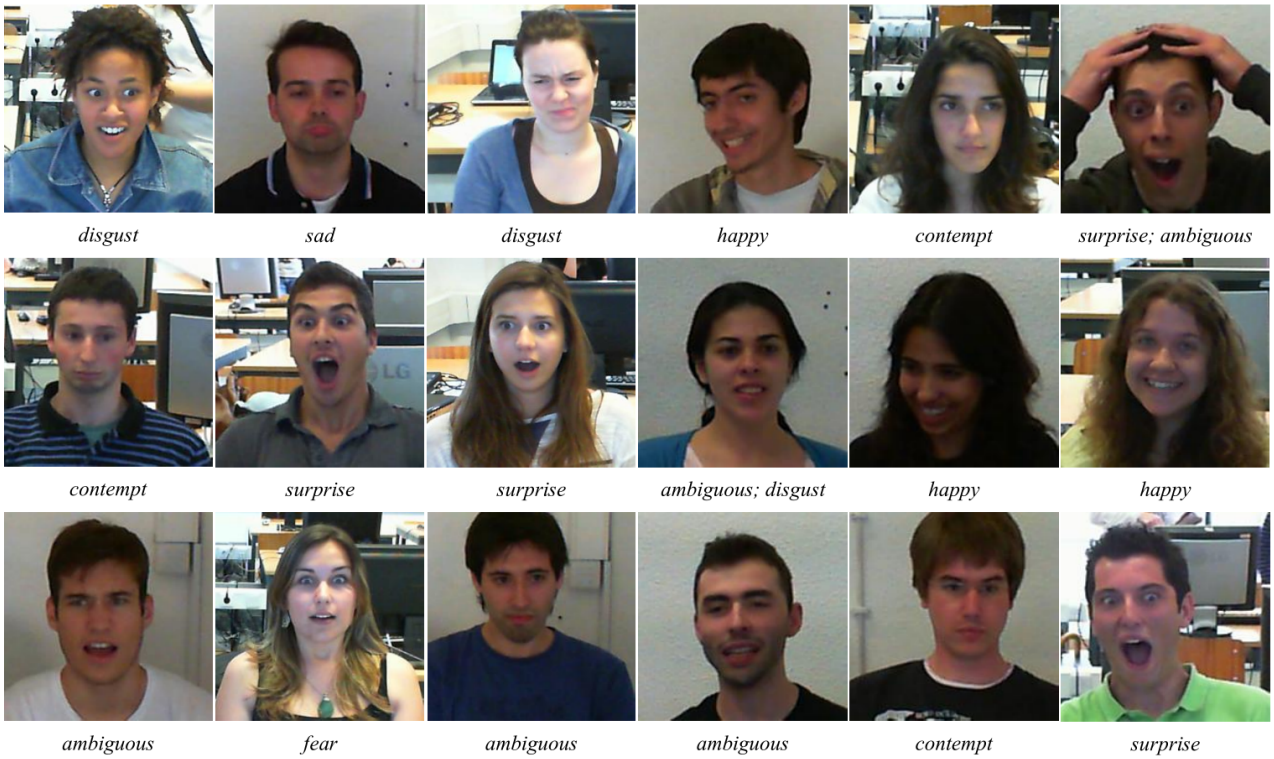


Figure 2: Example faces from the dataset.

job, whereas large maximum values may distract the worker after some micro-tasks and produce worse results. We limited each job to a maximum of 500 images.

4.2 Job attributes

Beyond worker qualification, there are some job attributes that we parameterized. The first three attributes directly change the job's cost.

Number of micro-tasks per page. A job is divided into several pages and we can define how many micro-tasks each page has. Usually a worker completes at least one page and as such, this parameter works as a minimum of judgments a worker must complete.

Price per micro-task. The price to pay for each micro-task can make the job cost increase significantly without increasing the results' quality. We observed that this parameter had no or little effect on the agreement quality.

Judgments per micro-task. A micro-task covers a set of images that must receive an given number of judgments. The larger the number judgments, the greater the confidence of our task design. We collected 5 judgments per item in each job.

Gold questions. The intentions and knowledge of each worker must be validated. For that, one must provide gold questions. In our micro-tasks gold questions are images with known facial expressions. Every worker must answer, at least, 4 gold questions before any other micro-task.

4.3 Worker interface

We conceived an interface that allows the worker to choose, for each image, the facial expression that best describes it (from a given set of choices) and the inherent intensity. The worker interface is presented in Figure 3. In this interface the worker must (1) explicitly select the player's facial **expression**, and (2) select the **intensity** of the facial expression. This second action is intended to disambiguate and to quantify the certainty that a worker assigns to its label.

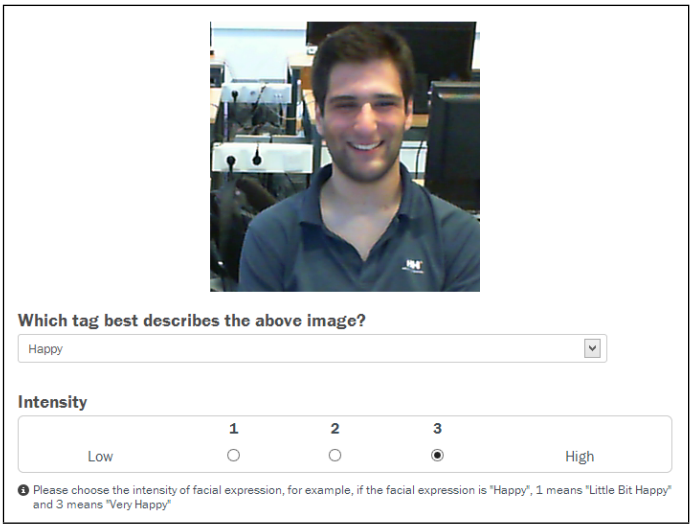


Figure 3: Worker interface.

5. RESULTS AND DISCUSSION

In this section we present the results collected through crowdsourcing for a sample of 500 images, randomly collected from our dataset with over 40,000 images. We ran 7 jobs and took into account all votes produced by all trusted workers in every job, making a total of 40 votes per image and 228 workers. The confusion matrix for the winner facial expression versus the other votes is presented in Table 1. The facial expression of an image is determined by the most voted tag. To compare the produced results, we take as examples the facial expressions with lowest and highest agreement ilustrated in Table 2.

High-agreement expressions. The highest agreement in our dataset is 1, which means that all 40 workers voted in the same facial expression, this is 10 % of the images in our dataset, and almost 50 % have agreement an above 0.8.

Analysis of facial-expression labels confusion matrix. Table 1 illustrates the judgments confusion among all six basic expressions, the composed expression *contempt*, an ambiguous and a noisy capture. The diagonal of the confusion matrix illustrates how the majority of expressions are clearly separable from the others (apart from the *fear* label, all expressions reached an agreement above 50%). The facial expression *happy* was the most consensual among all workers. Sometimes it is confused with *Neutral* due the intensity of the facial expression. One worker may consider a person grinning as *happy* while another worker may consider just *neutral*. The most dubious facial expression is *fear* which is often confused with *neutral*, once more due the intensity of expression.

Low-agreement, weak-labels. The facial expression images with low agreement exploits the advantage of *not* using the binary judgment model. The researcher is given the freedom to decide how to handle the label data. For example, one may want to use a weak-label approach by considering the most voted labels and the distribution of the votes.

Low-agreement, good counterexample. Although the dataset has images with low agreement, this makes them a good counter-example for some facial expressions. Regarding table 2 one may conclude that the facial expression on the left is *not* surprise nor happy.

Per-expression analysis. Figure 4 illustrates the distribution of workers agreement over each facial expression. It is interesting to observe the shape of these curves - ideally they should all start and end with an agreement of 1.0 (meaning that all workers agreed on one single label for every image). The area underneath the curve indicates the overall labeling agreement across all workers for that expression. The agreement curves for *sad*, *angry*, *fear*, *disgust* and *surprise*, show that some workers had an agreement of 0.0, which means that these workers failed all images for this expression. The exception to this trend occurs for the expression *happy* where the worst worker agreement was near 0.4. The shape of graph for the facial expression *Fear*, presented in Figure 4, is due to the existence of few images with facial expression *fear*.

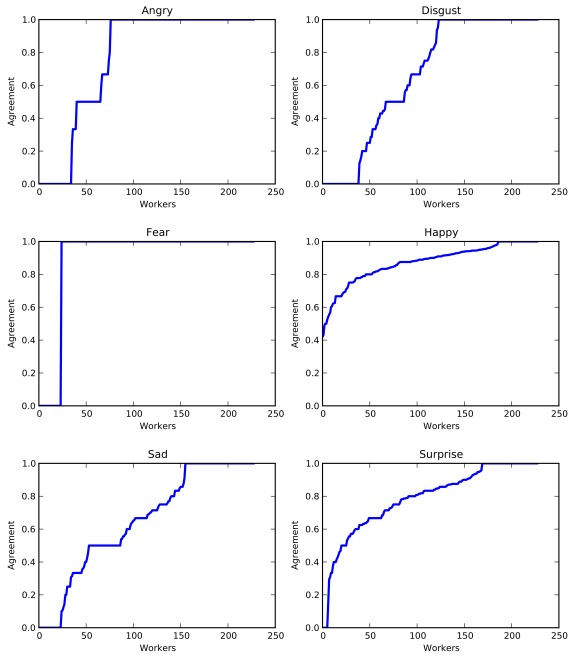


Figure 4: Workers agreement with the selected label, sorted by agreement

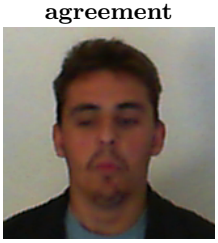
Expression	Lowest agreement	Highest agreement
		
Neutral	9	0
Angry	1	0
Disgust	9	0
Fear	1	0
Happy	0	40
Sad	5	0
Surprise	0	0
Contempt	0	0
Ambiguous	10	0
Not a face	0	0

Table 2: Workers' votes for facial expressions with lowest and highest agreement.

6. CONCLUSION

This article describes the crowdsourcing job design for tagging affective-interaction gaming data. In this paper we described the settings we examined to determine the best crowdsource job settings to maximize the agreement across workers' judgments. As a result, we release a dataset[3] with over 40,000 images of player's facial expression and multi-

[3]http://novasearch.org/datasets/

	Neutral	Angry	Disgust	Fear	Happy	Sad	Surprise	Ambiguous
Neutral	**60.1** %	4.6 %	5.5 %	25.7 %	3.5 %	9.4 %	2.3 %	10.5 %
Angry	3.4 %	**51.7** %	5.9 %	0.0 %	0.6 %	3.0 %	0.9 %	4.7 %
Disgust	3.8 %	8.3 %	**57.1** %	8.6 %	1.6 %	5.5 %	2.0 %	16.5 %
Fear	0.9 %	3.0 %	4.9 %	**31.4** %	0.4 %	2.7 %	3.5 %	1.3 %
Happy	6.8 %	9.6 %	2.4 %	2.9 %	**87.2** %	2.3 %	6.5 %	6.8 %
Sad	6.3 %	5.0 %	9.5 %	11.4 %	0.6 %	**63.2** %	0.7 %	1.8 %
Surprise	2.3 %	4.3 %	3.2 %	5.7 %	2.3 %	2.3 %	**79.7** %	7.9 %
Ambiguous	8.4 %	7.9 %	4.6 %	8.6 %	2.4 %	6.2 %	3.4 %	**42.4** %
Contempt	7.5 %	5.0 %	7.0 %	5.7 %	1.3 %	5.3 %	0.8 %	6.3 %
Not a face	0.5 %	0.7 %	0.0 %	0.0 %	0.1 %	0.1 %	0.2 %	1.8 %
Total	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %

Table 1: Confusion matrix for each facial expression

ple judgments per facial expression. Judgments for the full set of images are also provided to foster the investigation of other relevance models for affective-interaction.

7. REFERENCES

[1] S. Ahern, M. Davis, D. Eckles, S. King, M. Naaman, R. Nair, M. Spasojevic, and J. Yang. Zonetag: Designing context-aware mobile media capture to increase participation. In *Proceedings of the Pervasive Image Capture and Sharing, 8th Int. Conf. on Ubiquitous Computing, California*, 2006.

[2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980. ACM, 2007.

[3] anonymous. anonoymous.

[4] anonymous. anonoymous. In *ACM Multimedia 2013, accepted*, 2013.

[5] C. Darwin. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[6] P. Ekman. Facial expression and emotion. 48(4):384–392, 1993.

[7] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[8] A. Foncubierta Rodríguez and H. Müller. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, CrowdMM '12, pages 9–14, New York, NY, USA, 2012. ACM.

[9] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101, 2010.

[10] J. Moehrmann and G. Heidemann. Efficient annotation of image data sets for computer vision applications. In *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, VIGTA '12, pages 2:1–2:6, New York, NY, USA, 2012. ACM.

[11] 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 international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA, 2010. ACM.

[12] D. Roman. Crowdsourcing and the question of expertise. *Commun. ACM*, 52(12):12–12, Dec. 2009.

[13] N. Sawant, J. Li, and J. Z. Wang. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools Appl.*, 51(1):213–246, Jan. 2011.

[14] Y.-L. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In *Handbook of Face Recognition*, pages 247–275. Springer New York, 2005.

[15] T. Volkmer, J. A. Thom, and S. M. Tahaghoghi. Modeling human judgment of digital imagery for multimedia retrieval. *Multimedia, IEEE Transactions on*, 9(5):967–974, 2007.

[16] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[17] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, and J. Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. 2010.

[18] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–6, 2008.

[19] M.-C. Yuen, I. King, and K.-S. Leung. A survey of crowdsourcing systems. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 766–773, 2011.