Guess What? A Game for Affective Annotation of Video Using Crowd Sourcing

Laurel D. Riek[1], Maria F. O'Connor[2], and Peter Robinson[2]

[1] Department of Computer Science and Engineering, University of Notre Dame, USA
[2] Computer Laboratory, University of Cambridge, UK
lriek@cse.nd.edu

Abstract. One of the most time consuming and laborious problems facing researchers in Affective Computing is annotation of data, particularly with the recent adoption of multimodal data. Other fields, such as Computer Vision, Language Processing and Information Retrieval have successfully used crowd sourcing (or human computation) games to label their data sets. Inspired by their work, we have developed a Facebook game called *Guess What?* for labeling multimodal, affective video data. This paper describes the game and an initial evaluation of it for social context labeling. In our experiment, 33 participants used the game to label 154 video/question pairs over the course of a few days, and their overall inter-rater reliability was good (Krippendorff's $\alpha = .70$). We believe this game will be a useful resource for other researchers and ultimately plan to make *Guess What?* open source and available to anyone who is interested.

Keywords: social context annotation, emotion annotation, video annotation, human computation, crowd sourcing.

1 Introduction

One of the most substantial problems researchers in Affective Computing face is data labeling. Beyond the length of time the standard video production process takes (collecting, segmenting, converting), the labeling process is one of the most time-consuming and expensive parts of the research lifecycle [24].

With the recent move as a community toward using multimodal data [6,17,20], labeling time is further increased. Researchers may wish to label several aspects of activity, such as facial expressions, gesture, posture, speech, and prosody; as well as more holistic attributes, such as overall mood, social roles, situational context, and social norms [19].

In the Computer Vision community, von Ahn [1] pioneered the field of Human Computation (HC), or crowd sourcing, to help with image data labeling. The premise of HC games is to have thousands of non-experts play a fun game while unwittingly labeling large corpora of data. For image labeling, HC is a well-validated technique, and yields results comparable to those of trained expert labelers [15].

S. D´Mello et al. (Eds.): ACII 2011, Part I, LNCS 6974, pp. 277–285, 2011.

This approach has also been effectively used in Human Language Technology for textual data labeling [11,22], in Information Retrieval for improving search [14,7], and in Speech for prosody labeling [23].

Many of these efforts have yielded similarly positive, comparable results to those found in the Computer Vision literature regarding the efficacy of non-expert labelers. Hsueh et al. [11] suggest that for some labeling tasks, such as a sentiment analysis, even if the HC data is noisy it still provides very useful data for modeling. This is also the consensus of Sheng et al. [21] for Data Mining tasks.

Inspired by these findings, we developed an HC game called *Guess What?* for labeling multimodal, affective video data. This paper describes an overview of the game, in terms of its implementation details, scoring mechanism, and game play. We also describe a pilot evaluation of the game, where we use it to macro-label social context in amateur You Tube videos.

We believe this game will be a useful resource to other researchers and ultimately plan to make *Guess What?* open source and available to anyone who is interested.

2 The Game

2.1 Overview

Guess What? is a Facebook game in which players are shown a video clip and are then asked a question about it. The objective of the game is to earn points. A player earns a small number of points for each question answered, but in order to do well, they need to guess what answer most other people would give.

Figure 1 shows some example screen shots from the game for macro-labeling the social context of scenes, which we used during our initial evaluation of the game.

2.2 Game Play and Scoring

Players first receive a neutral audio/visual test to make sure their system is properly configured. Following this, they can either play a round of the game or view the high scores list.

In a round of the game, the player will will be shown a video (which can be of any length), and either a fixed-choice or open-response question. *Guess What* automatically adjusts its layout and scoring mechanism to accommodate whatever mix of video and question types the researcher specifies in their initialization.

After players answer a question, they are given a score. Two scoring mechanisms are used in the game. If a question is fixed-choice, the player is awarded points based on the percentage of other players who chose the same answer as them. For open-response questions, players receive 25 points for a new answer which the system has never been seen before, 75 points for other answers, and 100 points for the answer which has been chosen most often. Both scoring mechanisms favour players who consistently answer "correctly" as judged by all other users.

Fig. 1. Some sample screen shots from the *Guess What?* game. These are macro-level questions for labelling social context, but the game is easily configurable to also allow for micro-level annotation of video.

2.3 Implementation Details

As a crowd sourcing game, *Guess What* needs to reach the largest audience possible. This suggests two technical requirements - the game should run with minimal user effort (i.e., no downloading of plug-ins) and should be scalable to a large number of users.

For these reasons, *Guess What* is built using a java servelet backend with an HTML/Javascript user interface running on Google App Engine. App Engine is a request-driven, cloud-based application engine. As the application only consumes resources when required by a user, the cost of scaling is linear. A general overview of the server architecture is visible in Figure 2.

Google App Engine provides an object datastore, rather than a traditional relational database. All Videos and Questions are uploaded here. Researchers first need to provide a list of video URLs to App Engine, then use a web interface to run a servlet which reads this file and creates a blank datastore filled with video objects created using these URLs. Questions are also uploaded via a web interface.

When deciding what video a user will see, *Guess What* queries the datastore for the set of least-labelled videos coupled with ones the user has never seen. It then randomly selects one of these. This ensures both that a user who plays for

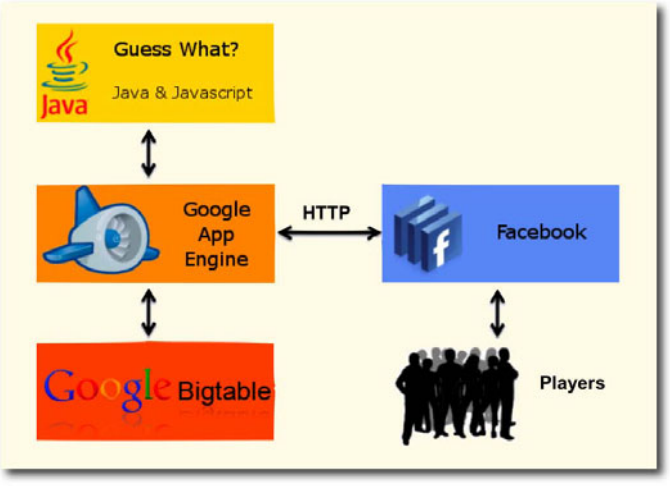

Fig. 2. A general overview of the game's server architecture

a long time will not see a repeating sequence of videos, and that labels will be assigned uniformly to the entire dataset.

For our initial deployment, we opted to deploy the game as a Facebook app. This was so we could utilize the Facebook authentication mechanism, as well as to allow people to easily share the game with their friends. While this was a successful strategy for the pilot version of our game, we found many users uncomfortable with Facebook's privacy settings for apps, so for our next deployment of the game we will switch to something else.

3 Evaluation

3.1 Data

In other work, we are investigating machine learning algorithms that can detect social context in naturalistic multimodal video data, in order to aid affective inference [16,19]. Therefore, when evaluating *Guess What?*, we opted to use data that will also help generate tags for our training set.

To generate an initial corpus of data to use in our evaluation, we searched You Tube for a variety of easily classifiable social events, such as birthday parties, weddings, sporting events, etc. We looked for videos with varying lighting conditions, camera angles, and quality, as we want our algorithms to be able to deal with as naturalistic data as possible. Because we were creating a multimodal corpus, we also searched for videos containing people speaking in different languages, playing different styles and kinds of music, and especially looked for people from a variety of ethnic, racial, and cultural backgrounds. Finally, we selected videos that were approximately three minutes in length.

Fig. 3. Some sample video clips of social scenes used as stimuli in the game evaluation. The videos varied in the number of people, social context, and activity, as well as the video quality, colors, and lighting.

We ultimately selected 39 videos that fit these criteria. The videos contained events such as concerts, sporting events, interviews, dinners, lectures, birthday parties, etc. A sample of clips can be seen in Figure 3.

3.2 Labeling Pilot

In order to establish fixed-choice labels for our main experiment, we conducted a pilot study using the *Guess What?* game.

Participants were recruited via word-of-mouth, and were sampled from the same population as our main experiment. Six people participated in the pilot, three female and three male, and their ages ranged from 21 to 34. All pilot participants were native English speakers.

In the pilot, after giving consent to participate and undergoing a neutral audio/visual test, participants were shown the 39 stimuli videos in random order. After watching each video, participants were asked seven open-ended questions about their social context.

We developed these questions based on definitions of social context suggested by Philippot et al. [18] and Burke and Young [5]. The questions were:

- Do these people know one another?
- What type of event is this?
- How formal is this event?
- What are these people doing?
- What's the overall mood?
- What type of people are these?
- What time is it?

Following the pilot, we examined the responses per video and per question and selected the annotations with 66% or greater agreement among raters. To resolve

Table 1. The four social context questions used in our main experiment, and the choice of responses generated from the pilot study

Question	Labels
What is the predominant activity in this video?	Cake Cutting, Celebrating, Cheering, Dancing, Eating, Getting Married, Gift Giving, Joking, Playing, Singing, Talking, Watching
What word best describes this event?	Birthday, Clubbing, Dinner, Interview, Lecture, Party, Performance, Sports, Wedding
What would you estimate the time of day to be?	Morning, Afternoon, Evening, Night, Unknown
What kind of occasion is this?	Formal, Informal

emotion-related synonyms, we used the 24 emotion groups from the Baron-Cohen taxonomy [2,8]. For example, the labels "happy" and "joyful" would both be considered "happy". For other synonyms we used a thesaurus.

We decided to remove three of the questions from our main experiment, regarding subject relationship, person type, and overall mood. The relationship and person type questions caused some confusion among our pilot participants, particularly for some videos that depicted several different groups of people, for example at performances (e.g., the video shows both audience members and performers).

We removed the mood question because for nearly all the events we chose had a positive valence, so nearly all the pilot mood labels were synonyms of "Happy" (30/39 videos). The remaining nine videos were lectures and interviews, and the most common label given for them was "Serious". Thus, we did not see much point in collecting additional mood annotations for this dataset since they were already so well-labeled in the pilot.

We refined the remaining four questions to be more clear, and the final questions and labels used in the main experiment are shown in Table 1.

3.3 Main Experiment

For the main experiment, we recruited participants via Facebook and word-of-mouth. 33 people participated in our main experiment, 15 female and 17 male. Of those who chose to complete our optional demographics survey, ages ranged from 19 to 75, and all but one participant considered themselves fluent in English. Participants lived in the United Kingdom, United States, Ireland, Indonesia, and Germany.

After giving informed consent, participants took a neutral audio/visual test. They then had the option to give voluntary demographics information. Following

this, they were presented one of the 39 videos in random order, with one of the four fixed-choice questions randomly assigned. (See Figure 1 for example screen shots from the experiment.) After submitting their answer, they were told their score and had the option to play again.

Participants could play as many times as they wanted, some played only once and some played well over 60 times, but on average participants played 22.29 times (s.d. = 22.74).

3.4 Results

For this evaluation we had one primary measure of interest, which was overall inter-rater reliability. We used Krippendorff's α, which is viewed as more reliable than other reliability measures when there are more than two raters [10]. Also, it is a robust measure capable of dealing with incomplete data, which one would expect when from crowd sourcing, and multiple nominal category levels, which for this experiment we had for each question (See Table 1.)

We used the SPSS macro developed by Hayes [9] for computing α. For this macro we provided all of our raw data, both labeled and missing, for the 33 participants labeling a possible 156 video/question pairs (39 videos x 4 questions).

Krippendorff's α was .702, which indicates fairly good reliability [13]. Krippendorf and Hayes suggest for most use of their measure, values between .667 and up are acceptable. Since this is a far more conservative measure than Fleiss' κ, we are confident for this experiment that we have good inter-rater reliability.

4 Discussion and Future Work

In this paper, we introduced our crowd sourcing game *Guess What*, and described an initial evaluation of it for macro-level labeling social context in videos. Based on our results and the positive feedback we received from participants, we believe this game will be a useful resource for other researchers and plan to make it open source and available to anyone who is interested.

In the future, we plan to experiment with using *Guess What?* for micro-level labeling [3], as well as allowing more fine-grained control to researchers on how questions are presented. For example, if questions were tied to particular subsets of videos, the game could be used in conjunction with machine learning algorithms which first crowd-sourced the high-level classifications then switched to a set of more fine-grained questions when a particular confidence level was reached.

We also plan to conduct a comparison between using *Guess What?* and performing traditional in-person experiments, as well as a comparison between trained and naïve labelers. For crowd-sourced multimodal affective data this sort of deeper methodological exploration is timely, as this is an important topic not just to the Affective Computing community, but to many other research communities as well (c.f. Bernstein et. al [4], and Kazai and Lease [12]).

Finally, we plan to extend our pilot and use *Guess What?* to macro- and micro- label a variety of aspects of social context, such as social norms, social

roles, situational context, and cultural conventions [19]. Our first aim will be to determine broad social context by asking high-level questions relating to the setting, personal relationships, event and time. It may also be interesting to look at non-obvious data which can be collected using the game. For example, geolocation information could be used to compare answers from different cultural areas, and simple changes could allow the separate collection of labels associated with sound or vision only.

All of these extensions will help us to better label social context in video, which will ultimately be used to improve affective inference.

Acknowledgements. This work is supported by the Google Anita Borg Memorial Scholarship, the Neil Weisman Fund, and the Qualcomm Research Studentship in Computing.

We are grateful to James Neve for the substantial development effort he contributed to the original game. We would also like to thank Leszek Swirski and Heather Keith Freeman for their assistance.

References

1. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 319–326. ACM, New York (2004)
2. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.J.: Mind reading: the interactive guide to emotions (2004), `http://www.jkp.com/mindreading`
3. Bavelas, J., McGee, D., Phillips, B., Routledge, R.: Microanalysis of communication in psychotherapy. Human Systems 11(1), 47–66 (2000)
4. Bernstein, M., Chi, E., Chilton, L., Hartmann, B., Kittur, A., Miller, R.: Crowdsourcing and Human Computation: Systems, Studies and Platforms (May 2011), `http://crowdresearch.org/chi2011-workshop/` (last accessed April 15, 2011)
5. Burke, M.A., Young, P.: Norms, Customs, and Conventions. In: Benhabib, J., Bisin, A., Jackson, M. (eds.) Handbook for Social Economics. Elsevier, Amsterdam (2010)
6. Calvo, R., D'Mello, S.: Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications. IEEE Transactions on Affective Computing 1(1), 18–37 (2010)
7. Ganjisaffar, Y., Javanmardi, S., Lopes, C.: Leveraging crowdsourcing heuristics to improve search in wikipedia. In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, WikiSym 2009, pp. 27:1–27:2. ACM, New York (2009)
8. Golan, O., Baron-Cohen, S.: Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. Development and Psychopathology 18(02), 591–617 (2006)
9. Hayes, A.: SPSS, SAS, and Mplus Macros and Code (2011), `http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html` (last accessed April 15, 2011)
10. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures (2007)

11. Hsueh, P.Y., Melville, P., Sindhwani, V.: Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, HLT 2009, pp. 27–35. Association for Computational Linguistics (2009)
12. Kazai, G., Lease, M.: TREC 2011 Crowdsourcing Track (November 2011), https://sites.google.com/site/treccrowd2011 (last accessed April 15, 2011)
13. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage Publications, Thousand Oaks (2004)
14. McCreadie, R., Macdonald, C., Ounis, I.: Crowdsourcing a news query classification dataset. In: Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010), pp. 31–38 (2010)
15. Nowak, S., Rüger, S.: 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, pp. 557–566. ACM, New York (2010)
16. O'Connor, M.F.: Automatic Understanding of Social Scenes. Master's thesis, University of Cambridge (2011)
17. Pantic, M., Rothkrantz, L.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91(9), 1370–1390 (2003)
18. Philippot, P., Feldman, R., Coats, E.: The social context of nonverbal behavior. Cambridge Univ. Pr., Cambridge (1999)
19. Riek, L.D., Robinson, P.: Challenges and opportunities in building socially intelligent machines. IEEE Signal Processing (2011)
20. Scherer, K., Banziger, T., Roesch, E.: A blueprint for affective computing: A sourcebook. Oxford University Press, Oxford (2010)
21. Sheng, V., Provost, F., Ipeirotis, P.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 614–622. ACM, New York (2008)
22. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.: 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, pp. 254–263. Association for Computational Linguistics (2008)
23. Tarasov, A., Delany, S., Cullen, C.: Using crowdsourcing for labelling emotional speech assets. In: W3C workshop on Emotion ML (2010)
24. Yan, R., Yang, J., Hauptmann, A.: Automatically labeling video data using multi-class active learning. In: 9th IEEE International Conference on Computer Vision, pp. 516–523. IEEE, Los Alamitos (2003)