

# Crowdsourcing in Emotion Studies across Time and Culture

Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson  
University of Cambridge  
United Kingdom

mmam3@cam.ac.uk, tb346@cam.ac.uk, Peter.Robinson@cl.cam.ac.uk

## ABSTRACT

Crowdsourcing is becoming increasingly popular as a cheap and effective tool for multimedia annotation. However, the idea is not new, and can be traced back to Charles Darwin. He was interested in studying the universality of facial expressions in conveying emotions, thus he had to consider a global population. Access to different cultures allowed him to reach more general conclusions. In this paper, we highlight a few milestones in the history of the study of emotion that share the concepts of crowdsourcing. We first consider the study of posed photographs and then move to videos of natural expressions. We present our use of crowdsourcing to label a video corpus of natural expressions, and also to recreate one of Darwin's original emotion judgment experiments. This allows us to compare people's perception of emotional expressions in the 19th and 21st centuries, showing that it remains stable through both culture and time.

## Categories and Subject Descriptors

I.4 [Image Processing and Computer Vision]: Miscellaneous

## Keywords

Emotion recognition, Crowdsourcing, Multimedia annotation

## 1. INTRODUCTION

Crowdsourcing has emerged as a technique that makes use of the wisdom of the crowd to help solve problems by outsourcing tasks to a distributed group of people. Recently, crowdsourcing was shown to be cheap, fast and comparable to expert labels in different fields [7], especially when the crowd is diverse, independent and decentralised [8]. Having access to a wide and diverse population attracted researchers to use crowdsourcing for psychology and social science experiments, where the diversity of the participants is crucial. Individuals in Western, Educated, Industrialised, Rich, and Democratic (WEIRD) societies cannot be a representative sample of the rest of the world [5], leading to conclusions that do not generalise well. The availability of crowdsourcing, however, allows collecting data from diverse populations.

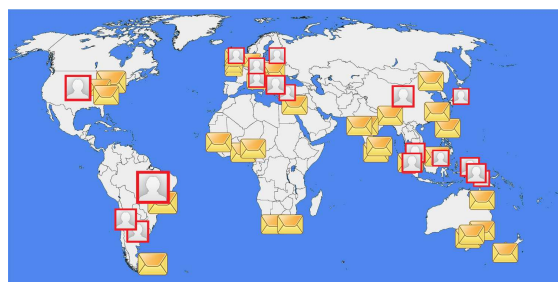


Figure 1: Map of Darwin's network (yellow letters) and Ekman's studies locations (Red squares), demonstrating the diversity of cultures explored.

## 2. EARLY CROWDSOURCING ATTEMPTS

Darwin was one of the first scientists to study emotions in a systematic way. Around 1866 he started to investigate the universality of expression and recognition of emotions [1]. This involved an early version of crowdsourcing - sending letters all over the world (See Figure 1), asking whether or not a particular expression was related to a particular emotion. This was, probably, the first crowdsourcing task in the history of scientific research on emotions. This allowed him to collect data from diverse sources. He even referred to it as his 'favoured research method'<sup>1</sup>

Furthermore, to study the recognition of emotions, Darwin used a judgement study on a set of photographs of facial expressions; a very modern approach at that time. He showed some Duchenne's photographs [2] of facial expressions of emotions to people to find out if they agreed about the emotion shown in each expression. Although Darwin's experiments lacked some of the rigour expected of psychology experiments today, his studies were cutting edge at the time; he used external observers and 'realistic' stimuli such as photographs. The main concept of his experiment is very similar to many contemporary studies on expressions and emotions. In Section 4 we describe a recreation of Darwin's experiment using 21st century crowdsourcing tools.

Ekman was inspired by Darwin [3], and set out to gather evidence in support of universality. The method that allowed him to convince the scientific community of universality of some facial expressions of emotion was the use of cross-cultural studies (See Figure 1), without which a con-

<sup>1</sup>We would like to thank Alison Pearn and Paul White at the Darwin correspondence project in Cambridge (<http://www.darwinproject.ac.uk/>) for providing original Darwin correspondences and transcript.

vincing argument could not have been made. Interestingly, the list of basic emotions that Ekman settled on [3] is very similar to the expressions used in Darwin’s experiments.

### 3. A RECENT CASE STUDY

With the advancement and widespread of the web, reaching to people in diverse locations is becoming much easier, and crowdsourcing is recognised as a useful tool for data annotation and labelling. We used community crowdsourcing to label complex emotions in Cam3D [6] videos of natural expressions.

Video segments were displayed randomly through a web interface and participants were asked to give one word describing the emotional state of the person in the video using free-form input, with an auto-complete list of mental states to avoid mis-spelling. The system was publicised via social networks and university mailing lists. As an incentive to participate, we offered gift vouchers to be won by five participants chosen through a raffle drawing of participants who labelled 50 videos or more. In the first run of the system, a total of 2,916 labels from 77 labellers were collected in 2 weeks ( $\mu=39$ ). After the system was featured by the BBC news, an additional 2,892 labels from 355 labellers were collected in one week ( $\mu=8$ ). Participants were not paid; they were rather motivated by the aim of the experiment in studying the correlation between expressions and emotions.

Out of the 451 segmented videos we wanted to extract the ones that can reliably be described as belonging to one of the 24 emotion groups from the Baron-Cohen taxonomy. For more details on the filtering see Mahmoud *et al.* [6].

We validated the labelling of the selected videos using Fleiss’s Kappa ( $\kappa$ ) [4] measure of inter-rater reliability. The resulting  $\kappa = 0.45$  indicates moderate agreement. This allows us to dismiss agreement by chance, and be confident in annotating the 108 segments with the emotional group chosen by the annotators. After performing the same pre-processing and statistical tests on second run data combined with first run, we ended up with Fleiss’s Kappa of 0.43 and 87 video segments where over 60% of subjects agreed on the emotional term. This is slightly lower agreement than the first run but still shows moderate agreement. Having a wider population makes it difficult to reach high agreement. However, reasonable agreement was still reached highlighting utility of crowdsourcing for emotion labeling.

### 4. MODERN RECREATION

We recreated Darwin’s original experiment using a crowdsourcing approach<sup>2</sup>. The goal was to study if the labels given to the photographs now would correspond to labels given by Darwin’s subjects. The experiment was performed using a similar web interface as the one described in Section 3. We displayed the photographs that were used by Darwin in his original experiments. The subjects for this experiment were anonymous and had no material incentive for task completion, besides interest in Darwin’s work. The task was quite short and took around two minutes on average. The experiment was publicised on the BBC news website, and acquired 202,303 labels in total, or 18,391 per photograph.

The results of our labelling experiment, compared to Darwin’s experiment, can be seen in Table 1. Agreement

<sup>2</sup>See <http://www.darwinproject.ac.uk/emotions>.

**Table 1: Comparing our recreation results [the most common labels chosen with their corresponding agreement rate] with Darwin labels with their corresponding agreement rates - based on his comments for the pictures (where available). Note the similarity of labels from 21st and 19th centuries.**

Our label	Agreement	Darwin label	Agreement
Happiness	37%	Laughter	88%
Surprise	91%	Surprise	96%
Fear	47%	Terror	83%
Sadness	82%	Grief and despair	-
Surprise	32%	Agony, torture & fright	69%
Confusion	31%	Crying from grief	60%
Sadness	32%	Suffering	-
Sadness	31%	Deep grief	93%
Surprise	44%	Fright with agony	80%
Boredom	43%	Hardness (surly reserve)	9%

for image labels was fair ( $\kappa = 0.28$ ). Our mean recognition rate for emotional images is 49%, which is comparable with similar studies of free rather than forced choice labels. All of the images had above chance performance when assigning an emotional term, indicating that the pictures used by Darwin indeed represent emotional states. Furthermore, 6 out of 10 pictures were given the same (or synonymous) label in both studies.

### 5. DISCUSSION

In the fields of psychology and social sciences, crowdsourcing can be considered an indispensable research tool. Although the term is relatively new, in studying emotions, ideas that shared properties of crowdsourcing existed since 1866, realising the importance of collecting data from large and diverse population in order to reach reliable conclusions that generalise across different cultures.

We discussed a recent case study of labelling a video corpus of natural expressions using a web interface. We have also compared emotion perception through time by recreating Darwin’s emotion experiment using community crowdsourcing through the same web interface. Our recreation experiment reinforces the belief that the ability to perceive emotions remains stable through time and cultures.

### 6. REFERENCES

- [1] C. Darwin. *The expression of the emotions in man and animals*. John Murray, London, UK, 1872.
- [2] G. Duchenne and R. Cuthbertson. *The mechanism of human facial expression*. Cambridge Univ Press, 1990.
- [3] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200, 1992.
- [4] J. Fleiss, B. Levin, and M. Paik. *Statistical Methods for Rates and Proportions*. Wiley, 2003.
- [5] J. Henrich, S. Heine, A. Norenzayan, et al. The weirdest people in the world. *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010.
- [6] M. Mahmoud, T. Baltrusaitis, P. Robinson, and L. Riek. 3D corpus of spontaneous complex mental states. In *Affective Computing and Intelligent Interaction (ACII)*, 2011.
- [7] R. Snow, B. O’Connor, D. Jurafsky, and A. Ng. Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks. In *Empirical Methods in Natural Language Processing*, pages 254–263, 2008.
- [8] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.