Investigating Crowdsourcing as a Method to Collect Emotion Labels for Images

Olga Korovina
University of Trento and
Tomsk Polytechnic University
olga.korovina@unitn.it

Marcos Baez
University of Trento and
Tomsk Polytechnic University
baez@disi.unitn.it

Fabio Casati
University of Trento and
Tomsk Polytechnic University
fabio.casati@unitn.it

Olga Berestneva
Tomsk Polytechnic University
ogb6@yandex.ru

Radoslaw Nielek
Polish-Japanese Academy of
Information Technology
nielek@pjwstk.edu.pl

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188667

Abstract
Labeling images is essential towards enabling the search and organization of digital media. This is true for both "factual", objective tags such as time, place and people, as well as for subjective, such as the emotion. Indeed, the ability to associate emotions to images is one of the key functionality most image analysis services today strive to provide. In this paper we study how emotion labels for images can be crowdsourced and uncover limitations of the approach commonly used to gather training data today, that of harvesting images and tags from social media.

Author Keywords
crowdsourcing; image tagging; emotions; subjective tasks

Introduction
Media tagging is an important tool for improving access to online resources [7]. Specifically, the ability to accurately label images with emotions is recognized as important for a variety of tasks [14, 12], and in particular for facilitating image search. Not surprisingly, most providers of AI services today offer tools for detecting and associating emotions to images (e.g. Google vision[1] and Microsoft Azure[2]), and emotion recognition apps are also flourishing.

[1] https://cloud.google.com/vision/
[2] https://azure.microsoft.com/en-us/services/cognitive-services/emotion/

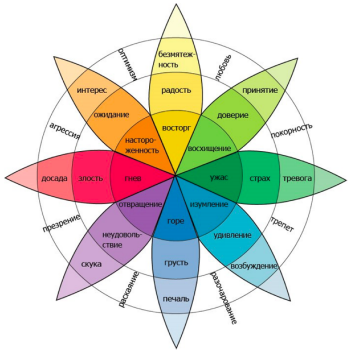

Figure 1: Plutchik Wheel of Emotions.

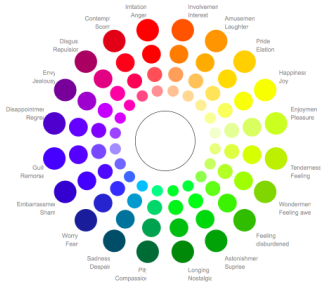

Figure 2: Geneva Emotion Wheel.

Machine Learning algorithms are usually trained with large datasets often obtained via crowdsourcing. This is also the case for emotion detection algorithms, where however crowdsourcing is often leveraged in a passive fashion: instead of asking crowd workers to label content, images are harvested from media libraries such as Instagram or Flickr based on associated tags [3] (see [17] for other sources of emotion-labeled datasets).

In this paper we explore some of the challenges in collecting a labeled dataset for emotion analysis. Specifically, i) we investigate *how subjective* emotional responses to images are, as high subjectivity would mean that searching for ground truths (as typically done) may not be the proper approach, ii) we investigate how robust emotion labeling is with respect to crowdsourcing task design, and iii) we assess whether the passive method often used so far actually provides labels that are consistent with how viewers respond emotionally to images.

While emotion labeling is an active field of research (including labeling via active crowdsourcing as discussed in [17]) and there are several datasets available, this set of problems has received little attention to date. Interesting research exists instead in the understanding of the different kinds of emotions we can experience and the relationships among emotions. This identification of emotions and their relationships is often conveyed in graphical form through "wheels", such as the Plutchick (Figure 1) and Geneva Emotion Wheel (Figure 2).

The *Plutchik Wheel of Emotions* [10] (PW) is a well-established psychological model of emotions used for structured tagging. The basic emotions (trust, disgust, surprise, anticipation, anger, fear, sadness and joy) are divided into opposite polarities (e.g., joy versus sadness) and each emotion has three degrees (e.g., serenity, joy, ecstasy). PW has been adopted in studies for commercial tagging [2] as well as in crowdsourcing contexts [9], and has been shown to be effective in initial studies also in terms of motivating respondents [12].

The Geneva Emotion Wheel (GEW) is similar but arranges information along the vertical and horizontal axis based on valence and dominance [13]. This representation has also been validated in theory and practice. Prior research also adopted the same method for labeling content other than images, such as speech [14].

These representations have been shown to be effective in collecting emotions in terms of commonly adopted metrics such as ability to collect a broad set of emotions, reduction of recourse to "other" categories, and the ease of understanding and tagging by non-experts [14]. However, they have not been studied in the context of crowdsourcing images, which requires, among other challenges, to screen malicious workers (something that is particularly difficult in subjective tasks) and to assess the ability of randomly selected users with possibly limited attention and motivation to provide a reliable dataset.

In this paper we report on a set of studies designed to assess the effect of different designs on how crowd workers associate emotions to images. Specifically, we design a process that is robust to cheaters while allowing for subjectivity and test two different methods for collecting labels (PW and GEW), assessing them in terms of commonly adopted measures for such tasks [14] such as agreement, emotion coverage, and simplicity. We also assess potential biases in the orientation of the wheel in the emotions collected. This is important as we know that task design in crowdsourcing can be influenced by a variety of aspects, not always easy to predict [16]. While identifying the "optimal" design is outside the scope of this paper, we

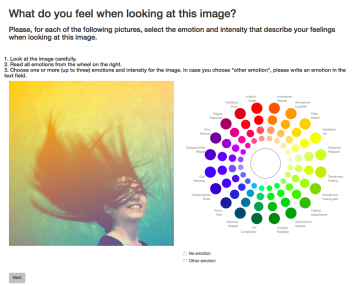

Figure 3: Image labeling task with GEW.

do want to understand if the collection method is robust to simple variations in the design.

Finally, we compare emotions collected based on PW and GEW with labels harvested from social media as commonly used by today's machine learning algorithms, and we show that labels collected via passive crowdsourcing are not consistent with those collected with an active process.

Method

To investigate the problem we design and run three sets of studies, devoted to i) assessing the properties of GEW and PW for crowdsourcing emotion labels as well as the degree of subjectivity in associating emotions to images, ii) assessing whether the specific wheel orientation affects the choices of workers (as one aspect of task design), and iii) comparing labels collected via active and passive crowdsourcing and discussing implications on using passive crowdsourcing as training data. Each of these points would probably require a set of studies on its own, so the goal of what follows is to identify if these are issues that are worth boht caution and deeper research.

The general task design is similar for all experiments. We first select a diverse (from emotion label perspective) set of images from top posts in social media based on image labels (passive crowdsourcing). In doing so, we look for labels that are also present in the wheels being evaluated. We then create crowdsourcing tasks where we propose images from this set to workers, in random order, along with one wheel and ask them to select up to three emotions by clicking on the corresponding part of the wheel (see Figure 3). Each participant would always use the same wheel to label the images. Respondents could also type in an emotion if the ones in the wheel did not correspond to how they felt, or could answer that no emotion was generated.

Additionally, workers answered two questions (based on a 5-item Likert-type scale), again, chosen based on the literature [14]: *It was easy for me to identify the observed emotions* and *The given method was sufficient to describe the observed emotions*.

We then implemented a simple process for detecting cheaters, a known and frequent problem in crowdsourcing [11]. Doing so is complicated by the fact that there is no "ground truth" for any item - something that also affects the ability to evaluate task design, so that the two issues are related. The crowdsourcing literature has been trying to devise various methods for filtering cheaters in this context, such as trying to detect if a worker clicks randomly [6], monitoring task execution time [4] and cursor trajectories [8], detecting outliers [15] or assessing consistency of answers [5]. Furthermore, besides subjectivity, emotion tagging is also affected by mood. For example, the lower the participant's mood, the more often they select sadness or calmness as emotions [1].

To address these issues we create a small set of test images on which we accept a broad set of emotions as valid responses. We do so on the basis on answers by trusted users and via small pilot crowdsourcing runs. We also assess consistency of responses on the same image. Notice that in this way we might filter out outliers along with cheaters, which we assume is acceptable or even desirable for typical applications but there may be scenarios for which this is not the case. The pilot also helped us uncover design issues and verify that users could understand the task and the questions. We ran the tasks on Crowdflower, in accordance with the guidelines for crowdsourcing in research[3].

[3](http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters)

Emotion name	Count
acceptance	35
admiration	55
aggressiveness	16
amazement	18
anger	28
anticipation	11
annoyance	21
apprehension	35
awe	23
boredom	17
contempt	33
disapproval	25
disgust	43
distraction	14
ecstasy	10
fear	25
grief	14
interest	60
joy	90
loathing	7
love	137
optimistm	56
pensiveness	13
rage	14
remorse	17
sadness	35
serenity	41
submission	13
surprise	32
terror	14
trust	29
vigilance	6

Table 1: Count of labels assigned using PW.

In a first experiment we aimed at assessing *agreement* among workers when labeling images as well as ease of labeling and coverage of emotions felt with PW and GEW. To do so we also borrow metrics and analysis from literature describing how this goal has been achieved for such wheels though for other kinds of content and not in crowdsourcing context [14]. We collected a small set of 32 pictures and asked each of 100 participants (50 for PW and GEW each) to label 12 of them (proposed in random order), two of which where tests. Participants were English-speaking, from UK, US or Canada.

A second experiment attempted to answer the question: *does the wheel orientation affect the participant's answers?*. In this experiment we have repeated the task three other times, with the exact same settings (task, images, and number of participants) as the previous one, but rotating the wheel by 90, 180 and 270 degrees. Rotation is only one, although important, aspect of task design (which also changes the positioning of emotions in the wheel with respect to the likely cursor placement achieved after clicking the "next" button).

The third experiment aimed at assessing the difference between emotions as collected via active and passive crowdsourcing. To do so we selected six representative emotions captured by both PW and GEW and collected 10 images for each emotion from both Instagram and Flickr, only tagged with that emotion. We collected, with the same method, 3 crowd label per image. Observing an effect in these experiment may be another indicator that developing training datasets for emotions is not something that can be taken lightly and requires research to identify emotions that are representative of what people feel.

Results

First, we examined differences between the two labeling methods. In general, PW seems to encourage selections of more emotions (average of 1.77 emotions per image in PW, 1.38 for GEW). A t-test confirms the difference as significant ($p < 0.0001$, t = 6.228). The difference is only partly explained by a higher selection of the *no emotion* answer in GEW (18% of responses) vs PW (10% or responses). In both cases there is a bias towards positive emotions: the most common emotions chosen were love, joy and interest in PW (Table 1) and happiness, enjoyment and involvement in GEW (Table 3).

We adopt Fleiss' kappa to measure (dis)agreement, as commonly done when there are more than two raters. Responses were considered as matching if they had at least one emotion among the ones they tagged in the same ray (leaf) of the wheel. We observed a Kappa value of 0.12 and a 0.14 for PW and GEW respectively, indicating slight agreement. The value is rather low despite the generous way to compute agreement, indicating that in general the subjective element is fairly strong in rating pictures (and is independent of the method adopted). We believe this observation makes it hard to consider a notion of ground truth, and indicate that caution is needed when filtering workers with test questions. We also investigated how labelers coped with the different labeling methods. The question "The given emotions were sufficient to describe the observed emotions" obtained a mean answer of 3.74 (sd=0.95, N=47) for PW and of 3.95 (sd=1.28, N=44) for GEW (the difference is not significant). We obtained similar results for the question "It was easy for me to identify the observed emotions" (mean=3.95, sd=1.28, N=44 for PW and mean=3.85,sd=0.50, N=48 for GEW).

Effect of wheel rotation. We analyze the effect of a rotated

Emotion name	Count
Amusement-Laughter	47
Astonishment-Surprise	9
Contempt-Scorn	18
Disappointment-Regreat	30
Disgust-Repulsion	30
Enjoyment-Pleasure	81
Envy-Jealousy	17
Embarrassment-Shame	7
Feeling disburdened-Relief	4
Guilt-Remorse	9
Happiness-Joy	117
Involvement-Interest	61
Irritation-Anger	21
Longing-Nostalgia	21
Tenderness-Feeling love	54
Pity-Compassion	18
Pride-Elation	29
Sadness-Despair	48
Wonderment-Feeling awe	24
Worry-Fear	32

Table 3: Count of labels assigned using GEW.

	Flickr	Instagram
joy	0.47	0.4
fear	0.47	0.27
surprise	0.03	0.07
disgust	0.17	0.53
anger	0.37	0.33
interest	0.23	0.23

Table 4: Proportion of active labeling that match passive assignments.

Rotation	Involvement	Tenderness	Pity	Disappointment	*Row Totals*
0	**335 (316.36) [1.10]**	112 (156.08) [12.45]	114 (107.59) [0.38]	116 (96.97) [3.73]	677
90	296 (323.37) [2.32]	191 (159.53) [6.21]	123 (109.97) [1.54]	**82 (99.12) [2.96]**	692
180	498 (515.43) [0.59]	282 (254.29) [3.02]	**172 (175.29) [0.06]**	151 (157.99) [0.31]	1103
270	450 (423.84) [1.61]	**194 (209.10) [1.09]**	128 (144.14) [1.81]	135 (129.92) [0.20]	907

Table 2: Chi-squared analysis of the effect of wheel rotation.

wheel by comparing the proportions of responses in each ray of the wheels with a chi-squared test. The result are shown in Table 2. In the figure, the column names are representative names for the wheel quadrants, and the cell contain the number of responses for emotions in that quadrant for each rotation, along with the expected cell totals and the chi statistics. The top right quadrant in each rotation is marked with bold font (the other proceed in clockwise fashion going towards right). The chi-square statistic is 39.376, and the result is significant at $p < .001$. While the data shows an effect of the rotation, it does not point to any specific quadrant, hinting and a combined influence of emotions and rotation and, consequently, at the need for studying the effect of task design in collecting emotion labels.

Consistency between active and passive crowdsourcing. We analyze results by computing the percentage of responses where the image label, as obtained from Instagram or Flickr, lies in the same GEW wheel ray (out of the 20 rays GEW has) of at least one of the 3 responses provided by the worker for the image. We call this a *hit*. Table 4 shows the percentages of hits, grouped by emotion. We can see that despite the conservative way of counting hits, the numbers are very low. Even if we remove "no emotion" responses and we count emotions on neighboring rays as half-hits (not shown), percentages only go up slightly.

Discussion. The main finding we do take home from the set of studies is that creating a dataset of emotion-labeled images can be prone to significant errors if done lightly. We saw that there is a high degree of subjectivity in the answer, pointing to the need of collecting fairly broad sets of emotions to be associated with images. We also saw that even a simple variation to task design such as orientation can change results, and that the common method of harvesting labels today does not correspond to what crowd worker would label as emotion they feel - possibly because passive labels reflect a context associated with a picture that workers do not have. Overall, we believe that the results, although preliminary, emphasize the need for research that considers these aspects in an area that is gaining so much importance in practice.

Acknowledgements

This project has received funding from the EU H2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 690962 and was performed by the authors in collaboration with Tomsk Polytechnic University within the project in Evaluation and enhancement of social, economic and emotional wellbeing of older adults, Agreement No.14.Z50.31.0029.

REFERENCES

1. Anna Aljanaki, Frans Wiering, and Remco C. Veltkamp. 2016. Studying emotion induced by music through a crowdsourcing game. *Information Processing and Management* 52, 1 (2016), 115–128. DOI: http://dx.doi.org/10.1016/j.ipm.2015.03.004
2. Dhanashri Chafale and Amit Pimpalkar. 2014. Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic. *International Journal of Computer Sciences and Engineering (IJCSE)* 2, 10 (2014), 14–18.

3. Vasavi Gajarla and Aditi Gupta. 2015. Emotion Detection and Sentiment Analysis of Images. *Georgia Institute of Technology* (2015).
4. Bruno Gardlo, Sebastian Egger, Michael Seufert, and Raimund Schatz. 2014. Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing. In *Communications (ICC), 2014 IEEE International Conference on*. IEEE, 1070–1075.
5. Tobias Hoßfeld, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. 2011. Quantification of YouTube QoE via crowdsourcing. In *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 494–499.
6. Sung-Hee Kim, Hyokun Yun, and Ji Soo Yi. 2012. How to filter out random clickers in a crowdsourcing-based study?. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization*. ACM, 15.
7. Ross McKerlich, Cindy Ives, and Rory McGreal. 2013. An Experimental Study of SocialTagging Behavior and Image Content. *International Review of Research in Open and Distance Learning* 14, 4 (2013), 90–103. DOI:http://dx.doi.org/10.1002/asi
8. Ricky KP Mok, Weichao Li, and Rocky KC Chang. 2015. Detecting low-quality crowdtesting workers. In *Quality of Service (IWQoS), 2015 IEEE 23rd International Symposium on*. IEEE, 201–206.
9. Radoslaw Nielek, Miroslaw Ciastek, and Wieslaw Kopec. 2017. Emotions make cities live. Towards mapping emotions of older adults on urban space. *arXiv preprint arXiv:1706.10063* (2017).
10. Robert Plutchik. 2001. The Nature of Emotions. *American scientist* 89 (2001), 344–350.
11. Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. 2011. Crowdsourcing subjective image quality evaluation. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 3097–3100.
12. Nina Runge. 2016. Tag Your Emotions : A Novel Mobile User Interface for Annotating Images with Emotions. (2016).
13. Klaus R Scherer. 2005. What are emotion? And how can they be measured? *Social Science Information Sur Les Sciences Sociales* 44, 4 (2005), 695–729. DOI: http://dx.doi.org/10.1177/0539018405058216
14. I. Siegert, R. Bock, B. Vlasenko, D. Philippou-Hubner, and A. Wendemuth. 2011. Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self assessment manikins. *Proceedings - IEEE International Conference on Multimedia and Expo* (2011), 0–5. DOI: http://dx.doi.org/10.1109/ICME.2011.6011929
15. Beatrice Valeri, Shady Elbassuoni, and Sihem Amer-Yahia. 2016. Crowdsourcing Reliable Ratings for Underexposed Items. In *Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST)*.
16. Jie Yang, Judith Redi, Gianluca DeMartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. In *Proceedings of The Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016)*. AAAI, 249–258.
17. Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark.. In *AAAI*. 308–314.