Correspondence

Creating Experts From the Crowd: Techniques for Finding Workers for Difficult Tasks

Luke Gottlieb, Gerald Friedland, *Senior Member, IEEE*, Jaeyoung Choi, Pascal Kelm, and Thomas Sikora, *Senior Member, IEEE*

***Abstract*—Crowdsourcing is currently used for a range of applications, either by exploiting unsolicited user-generated content, such as spontaneously annotated images, or by utilizing explicit crowdsourcing platforms such as Amazon Mechanical Turk to mass-outsource artificial-intelligence-type jobs. However, crowdsourcing is most often seen as the best option for tasks that do not require more of people than their uneducated intuition as a human being. This article describes our methods for identifying workers for crowdsourced tasks that are difficult for *both* machines and humans. It discusses the challenges we encountered in qualifying annotators and the steps we took to select the individuals most likely to do well at these tasks.**

***Index Terms*—Annotation, cheat detection, crowdsourcing, mechanical turk, multimodal location estimation, qualification.**

I. Introduction

CROWDSOURCED labor can be quite effective for implementing a task involving a high volume of data, as it allows the creator of the task to employ many workers for a short period of time. Traditional in-house annotation, for example, usually takes much longer, as it is limited by the number of people available locally to perform the task and by the facilities required to accommodate them. Crowdsourcing can be a cost-effective solution, but for the tasks described in this paper, the main advantage of crowdsourcing was increasing the speed annotations were generated. This article expands on our previous work [1], presenting the methods we used to find skilled workers on the Amazon Mechanical Turk crowdsourcing platform (MT) for an annotation task, comparing the output of the crowd workers with that of expert non-crowdsourced workers.

The specific task for which we first developed these methods, which we use as our main case study in this article, was to determine the geo-location of social media videos. Using crowdsourcing we can find a range of people who are close to being local experts for different areas, and then find the subset of those people who are talented at finding objective clues, such as street names or landmarks, in the data. We show that, when this type of knowledge and these skills are combined, crowdsourcing can be an important method for solving problems that would be impossible for individual workers.

This setup is significantly different from the standard tasks presented to crowdsourced workers, in that the standard tasks are usually relatively trivial for humans to perform, while being quite difficult for machines. The difficulty of the geo-location task presented challenges in the selection of Mechanical Turk participants, such as finding workers who would perform the task honestly. The qualification was designed to be challenging enough to qualify people for the real task while still being solvable by an individual in any location. Section II, shows similar work in crowdsourced task development. In Section III, we describe how we went about selecting the videos for the qualification task and the decisions that we made in designing the annotation interface.

Section IV describes how we approached qualifying crowd workers for a complex annotation task, including our initial trial run and the changes we made to our approach to improve the reliability of our results, and provides a summary of some key factors for developing the best labor pool for a given task. Section V describes the main phase of the project, collecting a human baseline for comparison with our automatic geo-location system, and briefly compares the efforts of our trained Mechanical Turk workers to the performance of our automatic system, noting patterns in where humans or machines outperform one another.

In Section VI, we summarize our findings about how improving the quality of a crowdsourced labor pool and provide some best practices for crowdsourcing difficult tasks that should be applicable across domains.

II. Related Work

Crowdsourcing is currently used for a range of applications, either indirectly, by exploiting unsolicited user-generated content, such as spontaneously annotated images [2], or directly, by utilizing systematic crowdsourcing platforms, such as Amazon Mechanical Turk, to mass-outsource artificial intelligence–type jobs [3]. In addition, crowdsourcing is used for conducting surveys and for evaluating user interfaces [4], designs, and technical approaches, where the task creators want to involve a very large number of subjects.

The discussion in [5] suggests that standard Mechanical Turk tasks tend to involve explicit evaluation, with the specific tasks ranging in complexity from reviewing, voting, and tagging to actually building artifacts. In our case, the artifact being built through this series of explicit evaluations was a structured knowledge base. Our workers were asked to find objective clues within the videos as to their locations, but they had to make a decision about each video regardless—which would, in the end, be a subjective judgment. Dishonest workers are a significant problem, as shown in [6], showing importance of screening and qualifying Mechanical Turk workers providing evidence that workers will attempt to game the system for quick rewards. One of the goals of our qualification metric was therefore to remove dishonest workers.

In [7] we see a good summary of recent work. The crowdsourced research that is most similar to ours is [8] on optical character recognition (OCR). This task is complex, and requires a degree of skill analyzing written language. The authors of "CrowdForge" [9] discuss the use of Mechanical Turk for complicated tasks, but their approach is quite different: breaking the complex task down into simple subtasks that humans can easily digest, and then intelligently combining them.

Manuscript received March 07, 2014; accepted June 03, 2014. Date of publication August 13, 2014; date of current version October 13, 2014. This work was supported in part by NSF EAGER under Grant IIS-1128599 and KFAS under a Doctoral Study Abroad Fellowship. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sheng-Wei (Kuan-Ta) Chen.

L. Gottlieb, G. Friedland, and J. Choi are with the Audio and Multimedia Research Directive, International Computer Science Institute, Berkeley, CA 94704-1198 USA (e-mail: luke@icsi.berkeley.edu; fractor@icsi.berkeley.edu; jaeyoung@icsi.berkeley.edu).

P. Kelm and T. Sikora are with the University of Technology, Berlin, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2347268

The approach we used in our research to deal with complexity and the need for specialized knowledge and skills was to provide tutorials and to use prequalification processes and/or include gold-standards items in HITs. There has been some opposition to employing such an approach from researchers who believe that similar or even better results can be achieved purely relying on redundancy [10]. However, we have found that collecting high numbers of redundant data points was difficult for a task like ours, as it requires too many resources, and it is not *a priori* clear how much redundancy is necessary to obtain good results in the first place.

III. Setup for the Experiment

The preparation for our experiment had two major parts, the selection of videos for the tutorial and qualification task, and the design and deployment of the Mechanical Turk–based user interface.

A. Video Selection and Expert Annotation

To support this experiment, we used the dataset provided for the MediaEval 2010 Placing Task [11] data set. This data set is comprised of Creative Commons licensed Flickr videos, including metadata with user-annotated titles, tags, descriptions, and comments, with information about the video's uploaders. The videos in this dataset had a maximum duration of 90 seconds, and an average length of less than 50 seconds. To get a representative sampling from the data set, we selected videos at random.

From these videos, we selected an initial 40 for which in-house worker's guesses matched the geo-location in the original metadata. In the first version of our qualifying task (described in more detail in Section IV), we used random subsets of 10 of these 40 videos; however, for later versions of the qualifying task, we narrowed it down further to a set of ten videos we presented to all of the workers, for reasons described below.

B. Developing the Web Interface

The next step in our setup process was developing a user interface for the Amazon Mechanical Turk workers to use for the qualification task described in Section IV. We went through several rounds of internal testing and feedback to improve the usability of the tool. The initial pretests were conducted on a relatively ad hoc basis, with various iterations of the tool being demonstrated to researchers working on unrelated tasks, who provided us with feedback. We then performed two rounds of internal testing of major prototypes, each with a small number of volunteers (between five and ten) from our lab. Some of the feedback we received led us to believe the annotators would benefit from a tutorial; this is explained in more detail in Section IV. Instructions were presented, and a progress bar was provided to indicate how far along the worker was in the task. Each video began playing automatically as soon as the page was loaded; all of the videos we used were re-uploaded to YouTube without their metadata.[1] A Google Maps instance was placed to the right of the video. When the worker clicked on the map, a marker would appear at that location that could then be dragged around the map. The marker's position was automatically translated into latitude and longitude. A placename search form, with spelling-correction, was also provided.

IV. Qualifying the Workers

This section explains how we approached the question of how to qualify crowd workers for a complex annotation task and the changes we made to our approach to improve the reliability of our results, and provides a summary of the factors we discovered were key in developing the best labor pool for the task.

A. Our Initial Approach to Qualification

For the first version of our qualification task, we created four HITs using randomly selected subsets of ten out of the forty videos selected by our in-house annotator Although we planned to vary the experimental conditions in the eventual collection phase (described in Section V) by presenting workers with different combinations of modalities, for the qualification phase we used only one condition: we presented all of the potential annotators with both the audio and video streams but no metadata, as we wanted to give them a task of moderate difficulty.

Before posting the qualification task on Mechanical Turk, we first asked in-house volunteers to attempt it, as a "baseline baseline" of expectation to help us set a qualification threshold for potential workers on the actual task. To evaluate the performance of the annotators, we compared the geodesic distance between the ground truth coordinates and the coordinates of the location estimated by the workers, using the Haversine distance. This measure is calculated thus:

$$d = 2 \cdot r \cdot \arcsin\left(\sqrt{h}\right) \tag{1}$$

$$h = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\psi_2 - \psi_1}{2}\right) \tag{2}$$

where d is the distance between points 1 and 2 represented as latitude (ϕ_1, ϕ_2) and longitude (ψ_1, ψ_2) and r is the radius of the earth (in this case, we used the WGS-84 standard value of 6,378.137 km).

After conducting several rounds of internal tests, we then tried an initial run with crowdsourced workers on Mechanical Turk. In those trials, we discovered that in presenting potential annotators with random subsets of videos, we had not taken into account the fact that, with only ten videos per subset, the distribution of difficulty levels would not be even, thus skewing the qualification results. Therefore, while we could compare annotators' performance on a per-video basis, we could not identify a threshold for qualification that was valid across the randomized subsets.

While some of the Mechanical Turk workers did relatively well on this initial trial of the qualification task, over half seemed not to understand it and were giving up in frustration, taking guesses at random, as indicated by their performance on videos we considered the de facto gold standard—those for which our internal testers had a near perfect location-detection rate. Every HIT we created contained at least some of these relatively easy gold-standard videos, so even though we did not have a reliable qualification threshold due to variation in average difficulty, we were at least able to reject submissions that were wildly inaccurate across all ten videos in the HIT. In our performance analysis, we also analyzed the time workers took to find the locations of a set of videos and eliminated the outliers; in many cases the reported times were so low, that it was clear that these workers were attempting to speed through the task for the bounty of USD 0.25, without making a serious attempt. (In these cases, the total time spent on the HIT was generally less than half of the duration of the video they were supposed to be viewing.) Due to these factors, we significantly revised our approach.

B. The Revised Approach

First, we created a short tutorial page providing a walkthrough showing how to locate an example video that most of our workers, both internal and external, had done quite poorly at locating in the first

[1]However, to comply with the terms of the Creative Commons license, the title and author/uploader information was shown at the end of each video.

Fig. 1. Video still from the tutorial, with an image of the City Market Building. From this image, an exact geo-location is discoverable.

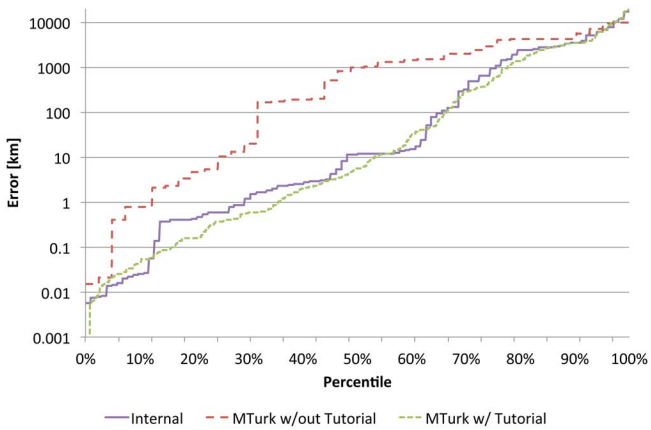

Fig. 2. Comparison of in-house workers and MT workers who had and had not received a tutorial, on the Ideal 10 video set.

iteration. Fig. 1 shows a frame that workers could use to determine the location where the video was filmed.

Second, we abandoned the use of randomized sets, based on our observations about the outliers in difficult sets in the first round. We selected a new qualification set made up of 10 videos (hereafter referred to as the "Ideal 10") where correct identification relied as little as possible on workers' previous experience of the location. Selecting videos for which identification was more likely to be based on interpretation of visual and audio clues than on instant recognition, and testing everyone on that same set, allowed us to identify a precise, reliable threshold for qualifying workers with good sleuthing skills, who would be able to deal with videos of unrecognized places in the actual data-collection phase.

Fig. 2 compares the performance of our internal testers, the MT workers in our initial qualification test, and the MT workers who had received the tutorial. The internal volunteer testers performed better than the Mechanical Turk workers, but with the addition of a tutorial and the honing of the test video set they identified the location of 52.8% of the videos to within 10 km, almost equaling the performance of the internal testers.

In Fig. 3, we see how accuracy of geo-location relates to the amount of time workers spent trying to classify a video. We found a similar correlation between time and accuracy across all of our videos. Table I shows the time workers spent to obtain different degrees of accuracy on videos in the qualification task. Workers who obtained an accuracy

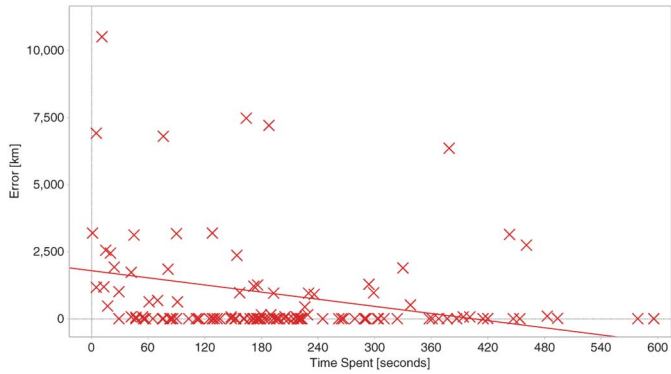

Fig. 3. Scatter plot of error versus time for a single video.

TABLE I
AVERAGE TIME SPENT ON INDIVIDUAL VIDEOS IN THE QUALIFICATION SET FOR RESULTS WITH A GIVEN MARGIN OF ERROR

1km	10km	50km	100km
88.70s	208.34s	177.58s	173.32s
500km	1000km	5000km	10000km
146.77s	139.15s	125.09s	119.97s

of within 10 km submitted their answers within 254 seconds, on average (min.: 10 secs, median: 181 secs, max.: 1966 secs). The average running time for videos in the Ideal 10 set was 52.8 seconds.

C. A Summary of Key Practices for Qualification

In sum, we found that the key ingredients to successfully qualifying a set of workers who could perform our annotation task were:

- Creating an informative tutorial;
- Implementing feedback methods so crowd-workers could influence the final design;
- Testing the qualification task with trusted workers to identify a realistic thresholds for performance;
- Ensuring that all potential workers received qualification HITs with the same difficulty level, to have a reliable basis for comparison;
- Monitoring the submitted qualification results for correlates of poor performance (such as overly-fast submission times).

V. COMPARING AN AUTOMATED SYSTEM WITH A HUMAN BASELINE

This section describes how we employed our qualified Mechanical Turk workers to create this baseline and briefly summarizes the results of the experiment enabled by our crowdsourcing efforts, these results are discussed in detail in Choi *et al.* 2013 [12].

A. Establishing the Human Baseline

We set the threshold for qualification of our Turk workers at 80% of answers being accurate to within 10 km of the ground truth geo-location for the video identified by the uploader. While the purpose of the experiment was to determine how accurately humans could perform this task on unknown videos, the qualification test was comprised of videos of known location. We chose the 80 qualified to do the actual task, yielding a pool of 290 skilled workers.

The interface we used for the data-collection stage was mostly identical to the one used in the qualification step, described in Section III-B, except that we varied the modalities presented (rather than presenting each video as audio+visual with no tags). We presented the media from least informative to most, i.e., first we gave the worker only the audio

stream, then the audio and visual together, and finally the audio and visual streams plus the textual metadata. In order to Each HIT included five videos with three different media combinations each, for a total of fifteen media streams. Each of the HITs was assigned to three different workers, allowing cross-validation. The main data collection phase took 18 days, we were able to collect a total of 11,900 localizations (2900 from qualification task, and 9000 from the experimental HITs).

The community on the website http://www.mturkforum.com/ was helpful and responsive, provided feedback on what was confusing in our experiment and on what motivated Mechanical Turk workers expect from HITs in terms of financial support, enjoyment, and author interaction.

B. Machine-Based Location Estimation

For the machine side of the comparison, we used multiple systems to deal with the different modalities in the data.

For audio-only location estimation, we used the city identification system described in [13]. We clustered the videos in the training set into the 40 cities they were recorded in or near, and tested the audio tracks extracted from the test videos against the trained models and picked the city with the highest likelihood, and converted the city labels produced by the automated system to the *(latitude, longitude)* format.

The system that analyzed the visual content of the videos for location estimation treated location estimation as an image retrieval problem, assuming that similar images mean similar locations—an approach similar to that of [14]. Our overall approach to multimedia-stream analysis is described in detail in [15]. To integrate textual and visual data, we combined the visual search method with the system described in [16].

While the soundtrack was still available to the humans in the second and third iterations they saw for each video, the machine baseline did not combine the three modalities (acoustic, visual, and text) in a single system because, as aforementioned, the audio-based location estimation system only covered a limited number of cities, while our visual- and text-based systems had broader coverage.

C. Results

Overall, humans performed better at this task than machines; 59% of the videos were located more accurately by the most accurate human than by the algorithm, when all modalities were utilized. However, this difference between the algorithm and (trained) human intelligence was relatively small, an 18% margin. In fact, in the very-accurate range (located to within 50 m of ground truth), our current algorithms outperformed humans by about a 12% margin in number of correctly located videos. On the other hand, when we compared our systems with the second-best of the three human results, the humans were more accurate than the algorithms for 40% of the videos. A brief analysis of the unqualified workers results on the qualification task indicates that machines would have dominated this task if qualification had not been used (The algorithms would have outperformed the humans around 80.

Both the machine and the humans performed better when information was available in more modalities. However, there were striking differences in the divergence between human and machine performance for different combinations of modalities. In the audio-and-visual case (using no textual metadata), humans performed better on 74.5% of the videos. Using only audio, the humans performed better on 83.6% of the videos. This latter disparity may be due in part to the fact that our acoustic-detection system produced only city-center coordinates (for the reasons described in Section V-B), while a human could choose any set of coordinates within a city.

VI. Conclusions

By changing our view on crowdsourced workers, we can use crowdsourcing to address tasks that are not straightforward and "mechanical". Rather than viewing the crowd as naive and allowing for their numbers to provide crowd wisdom, we can identify a subset of these workers who can be treated as de facto experts. In other words, depending on the task, it can be possible to find a subcrowd and train them to act as specialists for a given domain. In a sense, this approach creates an "elite crowd" that can be leveraged much as students are leveraged to perform research tasks at universities. The essential conclusion is thus that it is possible to use crowdsourcing for a very difficult task, but that one needs to have a systematic, well-planned procedure for selecting the members of the crowd who are most skilled at the task at hand.

The crowdsourcing approach that we developed for our geo-tagging work should be applicable to other tasks that require highly skilled crowdsourced workers. The following steps should be incorporated when designing such a task:

1) Ask experts with a strong understanding of the project requirements to perform the task, to identify a reasonable threshold of performance for qualifying crowd workers;
2) Conduct several rounds of testing with trusted workers, to refine the project design and obtain early feedback;
3) Perform a limited beta test of the application with crowdsourced workers, requesting feedback from those workers about anything they find confusing;
4) Develop a step-by-step tutorial showing the potential workers how they can succeed at the task.
5) Refine the tutorial and adjust the qualification task based on comparing the results from the crowd workers in the beta test to the results produced by the trusted expert workers, until the results are similar for the two groups.

Acknowledgment

Experiments involving human subjects were authorized under the Institutional Review Board of University of California, Berkeley, Berkeley, CA, USA, CPHS 2011-06-3325. The authors would like to thank H. Lei for his contribution to this work, and the users of http://www.mturkforum.com/ for their valuable insights.

References

[1] L. Gottlieb, J. Choi, P. Kelm, T. Sikora, and G. Friedland, "Pushing the limits of mechanical turk: Qualifying the crowd for video geo-location," in *Proc. ACM Multimedia 2012 Workshop Crowdsourcing Multimedia*, New York, NY, USA, 2012, pp. 23–28 [Online]. Available: http://doi.acm.org/10.1145/2390803.2390815

[2] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, pp. 157–173, 2008 [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0090-8

[3] P. G. Ipeirotis, "Analyzing the Amazon Mechanical Turk marketplace," *XRDS: Crossroads, ACM Mag. Students*, vol. 17, no. 2, pp. 16–21, Dec. 2010 [Online]. Available: http://doi.acm.org/10.1145/1869086.1869094

[4] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc. 26th Annu. SIGCHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2008, pp. 453–456 [Online]. Available: http://doi.acm.org/10.1145/1357054.1357127

[5] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Commun. ACM*, vol. 54, no. 4, pp. 86–96, Apr. 2011 [Online]. Available: http://doi.acm.org/10.1145/1924421.1924442

[6] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system?: Screening mechanical turk workers," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2010, pp. 2399–2402 [Online]. Available: http://doi.acm.org/10.1145/1753326.1753688

[7] M. Bernstein, E. H. Chi, L. Chilton, B. Hartmann, A. Kittur, and R. C. Miller, "Crowdsourcing and human computation: Systems, studies and platforms," in *Proc. 2011 Annu. Conf. Extended Abstracts Human Factors Comput. Syst.*, New York, NY, USA, 2011, pp. 53–56 [Online]. Available: http://doi.acm.org/10.1145/1979482.1979593

[8] G. Little and Y.-a. Sun, "Human OCR: Insights from a complex human computation process," 2011 [Online]. Available: http://crowdresearch.org/chi2011-workshop/papers/little.pdf

[9] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, "Crowdforge: Crowdsourcing complex work," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, New York, NY, USA, 2011, pp. 43–52 [Online]. Available: http://doi.acm.org/10.1145/2047196.2047202

[10] D. Karger, S. Oh, and D. Shah, "Budget-optimal crowdsourcing using low-rank matrix approximations," in *Proc. 2011 49th Annu. Allerton Conf., Commun., Control, Comput.*, Sep. 2011, pp. 284–291.

[11] M. Larson, M. Soleymani, P. Serdyukov, V. Murdock, and G. Jones, "Working notes," in *Proc. MediaEval 2010 Workshop*, Pisa, Italy, 2010.

[12] J. Choi, H. Lei, V. N. Ekambaram, P. Kelm, L. R. Gottlieb, T. Sikora, K. Ramchandran, and G. Friedland, "Human vs machine: Establishing a human baseline for multimodal location estimation," in *Proc. ACM Multimedia*, A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D. A. Shamma, M. Worring, and R. Zimmermann, Eds., 2013, pp. 867–876 [Online]. Available: http://dblp.uni-trier.de/db/conf/mm/mm2013.html#ChoiLEKGSRF13

[13] H. Lei, J. Choi, and G. Friedland, "City-Identification on Flickr Videos Using Acoustic Features," ICSI, Tech. Rep. TR-11-001, 2011.

[14] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. 2008 EEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[15] J. Choi, H. Lei, and G. Friedland, "The 2011 ICSI video location estimation system," in *Proc. MediaEval*, 2011.

[16] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran, "Multimodal location estimation of consumer media: Dealing with sparse training data," in *Proc. 2012 IEEE Int. Conf. Multimedia Expo (ICME)*, 2012, pp. 43–48.