

Pushing the Limits of Mechanical Turk: Qualifying the Crowd for Video Geo-Location

Luke Gottlieb, Jaeyoung Choi, Gerald
Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
{luke, jaeyoung,
fractor}@icsi.berkeley.edu

Pascal Kelm, Thomas Sikora
Communication Systems Group
Technische Universität Berlin
Sekt. EN1, Einsteinufer 17
10587 Berlin, Germany
{kelm, sikora}@nue.tu-berlin.de

ABSTRACT

In this article we review the methods we have developed for finding Mechanical Turk participants for the manual annotation of the geo-location of random videos from the web. We require high quality annotations for this project, as we are attempting to establish a human baseline for future comparison to machine systems. This task is different from a standard Mechanical Turk task in that it is difficult for *both* humans and machines, whereas a standard Mechanical Turk task is usually easy for humans and difficult or impossible for machines. This article discusses the varied difficulties we encountered while qualifying annotators and the steps that we took to select the individuals most likely to do well at our annotation task in the future.

Categories and Subject Descriptors

H5.2 [Information interfaces and presentation (e.g., HCI)]: User Interfaces: Evaluation/methodology—*Theory and Methods*; H5.3 [Information Interfaces]: Group and Organization Interfaces - Web-based interaction

General Terms

Design, Experimentation, Human Factors

Keywords

crowdsourcing, annotation, cheat detection, Mechanical Turk, qualification, multimodal

1. INTRODUCTION

The ubiquitous nature of the video camera, be it in camera phones or handheld devices has led to high numbers of quality geotagged videos on social networking sites, such as YouTube or Flickr. This represents a quantity of training

data on a vast and heretofore unprecedented scale. Therefore research projects have started exploring using this data to estimate the geo-location of a given video automatically, even when geo-location metadata isn't available. To aid this type of research, we have begun a project of collecting a "human baseline" for comparison with automatic systems and improvement of them, using Amazon's Mechanical Turk.

In this article we describe the methods we took for finding skilled Mechanical Turk participants for our annotation task, which will be to determine the geo-location of random videos from the web. The task itself is unlike the standard setup for a Mechanical Turk task, in that it is difficult for *both* humans and machines, whereas a standard Mechanical Turk task is usually easy for humans and difficult or impossible for machines. There are several notable challenges to finding skilled workers for this task: First, we must find what we termed "honest operators" i.e., people who will seriously attempt to do the task and not just click quickly through it to collect the bounty. Second, we need to develop meaningful qualification test set(s) that are challenging enough to allow us to qualify people for the real task, but were also solvable by individuals regardless of their culture or location, although English language understanding was required for instructions.

The paper is broken down into the following parts: In Section 2, we discuss of related or similar work in crowdsourcing, in Section 3, we examine the MediaEval 2010 Placing Task dataset and how we used it in this task. In Section 4 we explain how we selected the videos for the qualification task and present the Mechanical Turk user interface which we developed, then in Section 5, we describe our first qualification task, compare it to internal results, and examine why the results were poor. In Section 6, we describe the revised approach, the steps we took to improve the user experience, and then compare those results with our internal studies. In Section 7, we compare our Mechanical Turk results to those of a similar experiment which used a more traditional group of annotators, finally Section 8 concludes the article.

2. RELATED WORK

Crowdsourcing is currently used for a range of applications, e.g. exploiting unsolicited user contributions, such as spontaneous annotation of images for retrieval [7], or utilizing systematic crowdsourcing platforms, such as Amazon Mechanical Turk, to mass outsource artificial intelligence jobs [2]. Also, crowdsourcing is used for surveying and

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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM Multimedia '12 Nara, Japan

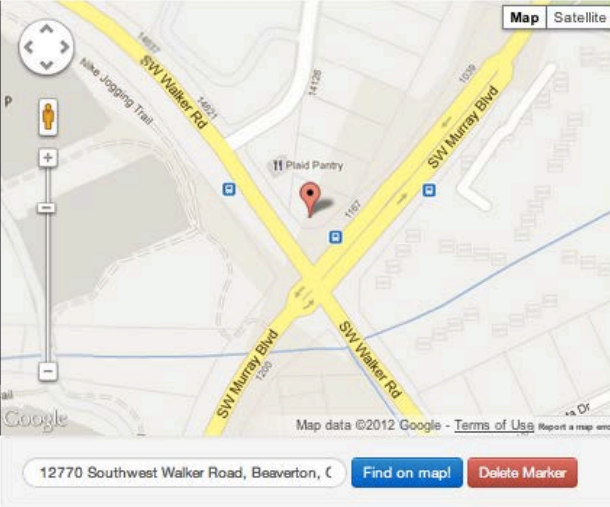

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

Estimate the Location of the Scene Presented in the Audio/Video.

Instructions Hide

- Use any resource available to you (including web search, encyclopedias, and personal knowledge) to estimate the location of the scene shown in the video as accurately as possible. Your answer should be in the format of GPS coordinate (latitude/longitude) of the location.
- You can obtain the coordinates with the aid of the map interface on the right. The map interface can be zoomed and dragged. The coordinate is determined by a mouse click on the location in the map which will automatically update your answer.
- You can use the 'Find on map!' text form to quickly jump to a desired location on the map using a textual search query, e.g. "Eiffel Tower, Paris".
- If your best estimate is on the scale of a city, a country or even a continent, you should place the crosshair in the center of the smallest estimated area (e.g. the center of a country).
- When you are finished with the current video, click the 'Submit' button to submit your answer and move on to the next video.

Tip If the map interface freezes, zooming in/out will fix it.



Latitude:

Longitude:

Figure 1: Screenshot of web interface used in the Amazon Mechanical Turk experiments described here.

evaluating user interfaces [4], designs, and other technical approaches so that subject numbers can grow very large. However, as the name accidentally implies, platforms such as Mechanical Turk are often best for mechanical tasks, i.e. tasks that do not require more than people's uneducated intuition as human being. Therefore, for a task like the one presented here, where humans and machines seem to perform very similarly and there is no clear intuition on how to solve the task, one has to be very careful about how to approach it properly. We believe, there is no previous work on using Mechanical Turk for geo-tagging videos, moreover there seems to be no previous work on how to use Mechanical Turk for a task that is not straightforward to solve. A good summary on recent work is provided in the proceedings of a SIGCHI workshop [1]. The work that comes closest to ours is [6] where OCR is crowdsourced. The task is already described as complex, yet OCR is clearly manageable by people who have learned to read. The approach described in our paper is to pre-qualify people, however, there is opposition to such an approach, for example by purely relying on

redundancy [3]. However, we believe, relying only on redundancy is unmanageable for a task like the one that we are describing because it takes too many resources and, most of all, it is not clear how redundant one has to be to obtain good results.

3. DATASET DESCRIPTION

All experiments described in this article were performed using the dataset distributed for the Placing Task of the MediaEval benchmark¹. The Placing task is part of the MediaEval benchmarking initiative and requires participants to assign geographical coordinates (latitude and longitude) to each provided test video. Participants can make use of metadata and audio and visual features as well as external resources, depending on the run.

The MediaEval Placing Task 2010 data set consists of Creative Common-licensed Flickr videos. The metadata for each video includes user-annotated title, tags, description,

¹<http://multimediaeval.org/>

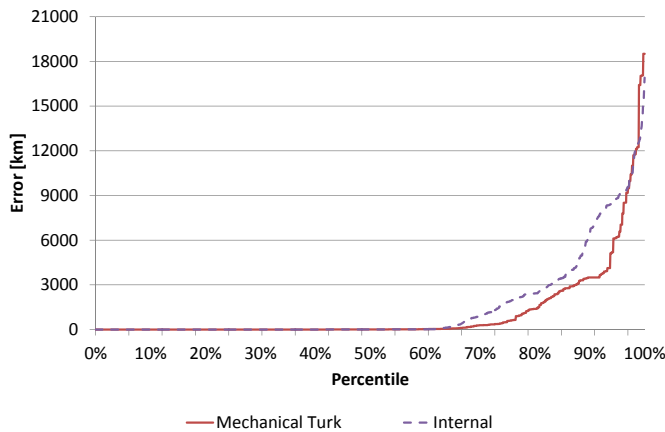


Figure 2: Initial comparison of internal workers and Mechanical Turk workers with 40 videos, as discussed in Section 5.

comments and also information about the user who uploaded the videos. Additionally, the metadata also include information about the user’s contacts, favorites, and all videos. The data set consists of 5091 training videos and 5125 test videos. For the qualification task, however, we only showed the videos (including audio).

According to [5], videos were selected both to provide a broad coverage of users, and also because they were geo-tagged with a high accuracy at the “street level”. Accuracy shows the zoom level the user used when placing the photo on the map. There are 16 zoom levels, and these correspond to 16 accuracy levels (e.g., “region level”, “city level”, “street level”). The relatively short lengths of each video should be noted as the maximum length of Flickr videos is limited to 90 seconds. Moreover, about 70% of videos in our data set have less than 50 seconds playtime. Flickr requires that an uploaded video must be created by its uploader. Manual inspection of the data set led us initially to conclude that many of visual/audio contents lack reasonable evidence to estimate the location without textual metadata. For example, some videos were recorded indoors or in a private space such as the backyard of a house. This indicates that the videos are not pre-filtered or pre-selected in any way to make the data set more relevant to the task, and are therefore likely representative of videos selected at random.

However, metadata provided by the user often provides direct and sensible clues for the task. 98.8% of videos in the training set were annotated by their uploaders with at least one title, tags, or description, often including location information. For a human, it is a fairly straightforward task to determine from the metadata which keyword or keywords combination indicates the smallest and most accurate geographical entity.

4. EXPERIMENTAL SETUP

4.1 Video Selection

The setup of this task had two important parts, the selection of videos and the design and deployment of the Mechanical Turk user interface. The task of video selection was relatively straightforward, although during the process of selecting videos we had to make several important decisions



Figure 3: Image of City Market Hall from tutorial. Humans should be able to find the location using a couple of web searches.

on the types of videos which were useful in the qualification task. In order to provide a representative sampling of the dataset, we randomized the complete list of videos, then our annotator viewed a subset of that list, and attempted to determine the location that the video presented. The annotator was allowed to use video and audio information, but not meta-tags, and was instructed to spend no more than 5 minutes per video. From this we collected the initial 40 videos which we used in our initial approach discussed in Section 5. Our discoveries there led us to take a subset of those videos, which we internally call the “Ideal 10” set, which we used in Section 6. In condensing the videos we tried to reduce the requirement for information from worker’s previous experience as much as possible, e.g. in the initial set there were videos of people in Machu Picchu, which our annotator immediately recognized, however there were no clues to reveal this location that would be usable to someone who had not heard or seen this location previously.

4.2 Development of Web Interface

The second component of our setup was the development of a user interface which the Amazon Mechanical Turk workers used for the Qualification task. In Figure 1, we provide an image of the current version of this interface. We went through several rounds of internal testing and feedback to enhance the usability of the tool. One of our more important discoveries was how the addition of a tutorial greatly aided the workers, as described in Section 6.

The instructions on the top of the screen can be expanded and shrunk with a ‘Show/Hide’ button. It was shrunk by default to make the whole interface fit in a normal-sized window to minimize unnecessary scrolling of the screen. A progress bar was shown below the instructions box to let workers know where they are along the progress of a HIT. A video was played automatically once the page was loaded. All Flickr videos were re-uploaded to YouTube without the metadata so that simply following the link on the player would not reveal any additional information about the video. To comply with the terms of the Creative Commons license, title of the video and the uploader’s information was shown at the end of each HIT to give credit to the original author.

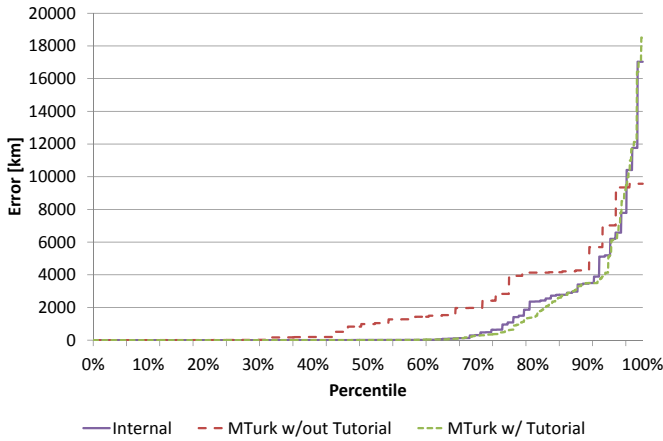


Figure 4: Comparison of performance results for Ideal10 set.

A Google Maps instance was placed to the right of the video. A marker would be dropped where the map was clicked, and it could be dragged around the map. The marker’s position was automatically translated to the latitude and longitude and printed to the ‘Latitude’ and ‘Longitude’ boxes. A location search form was placed under the map to aid the search of the location. The form had an auto-completion feature which would help in cases where the Worker did not know the exact spelling of the place, etc.

At the end of the HIT, we asked participants to leave comments about the HIT. We attempted to update the interface to reflect the feedback about the usability between our two approaches, and will be using the information that we received to make further improvements to the interface when we move on to the actual corpus collection section of our project.

4.3 Evaluation

To evaluate the performance of the online workers, the geodesic distance between the ground truth coordinates and those of the outputs from participants were compared. To take into account the geographic nature of the evaluation, the Haversine distance was used. This measure is calculated thus:

$$d = 2 \cdot r \cdot \arcsin(\sqrt{h}) \quad (1)$$

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

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, the WGS-84 standard value of 6,378.137km was used).

5. INITIAL APPROACH

In our initial approach to the qualification task we created four randomly selected subsets of our 40 videos selected by annotators. We then asked internal volunteers to attempt the task, as a “baseline baseline,” to give us some expectation of how well a Mechanical Turk worker might be able to perform, so that we could set a qualification threshold for potential workers on the actual task. After conducting

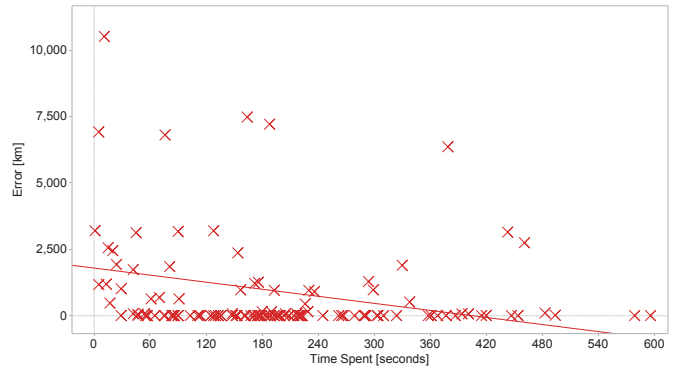


Figure 5: Scatter plot of accuracy (error) vs. time for a single video

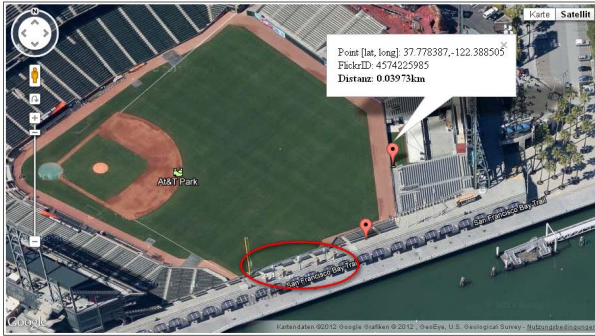
several rounds of internal tests and then experimenting on Mechanical Turk, we discovered that by making random sets we had not taken into account that the videos to be classified would be of varying degrees of difficulty, thus while we could compare the performance on a per video basis, we could not provide a threshold for qualification in the general sense. The bounty for this task (and all other tasks later described in this article) on Mechanical Turk was US\$ 0.25 per one HIT (10 videos). Figure 2 shows a comparison of the performance of our internal testers and the Mechanical Turk workers. While some of the Mechanical Turk workers did relatively well on the task, there were a significant number who seemed to not understand the parameters of the task, and were giving up in frustration, taking guesses at random. It is worth noting that we were also monitoring the time taken by the workers to locate a video; in our performance analysis we eliminated the outliers who were clearly attempting to speed through the task for the bounty, without making a legitimate attempt. We rejected submissions that were wildly inaccurate across all entries, as some of the videos in each set were quite easy and served as a gold standard training sets.

6. REVISED APPROACH

For our second round of qualification attempts we revised our approach based upon the results of the first attempt. We created a tutorial page that provides a walkthrough showing how to locate a video that most of the workers, both internal and external, did quite poorly on. Figure 3 shows the relevant frame that workers could use to determine the location the video was filmed. A Google image search of “City Market Hall” brings up a large number of images, and following some of those links will lead you to pages for the city of Roanoke, Virginia. A Further search of Google maps will give the exact latitude and longitude of the building in question.

As explained before, we abandoned the use of 4 randomized sets, and hand selected 10 videos (internally referred to as the “Ideal 10” set) for a new qualification test. Doing this would allow us to compare our qualification results on a set basis and have a more precise threshold for qualification.

In Figure 4 we have a comparison of the performance results for our internal testers, the initial test results, and the results with the tutorial. The internal tester and initial test results were derived by using the classification results from



Tags: dpm563, davidspointlessminute, sanfranciscogiants, battingpractice, maysfield, southpaw

Figure 6: This example shows when the human workers provided better annotation than the ground truth given by uploader of the video.

Table 1: Average time (in seconds) for a given margin of error

1km	10km	50km	100km
88.7s	208.34s	177.58s	173.32s
500km	1000km	5000km	10000km
146.77s	139.15s	125.09s	119.97s

the first round for the videos in the Ideal 10 set, as those videos are a subset of our larger annotation. We can see then that while the internal testers still perform better than the Mechanical Turk workers, the addition of a tutorial greatly narrowed the performance margin.

In figure 5 we see how accuracy of geolocation relates to time spent in the attempt to classify the video. We were able to find a similar correlation between time and accuracy across all of our videos, however it appears that in some cases the worker would spend a great deal of time on a video, while still getting a very poor result. While it is possible that in some cases this indicates the person is having a hard time on a particular video, often we will see that they spent a long period of time on all of the videos in their qualification set. In this case it is likely that they are pretending to do the work, rather than making an honest attempt, and we can reject their results. Workers who obtained under 10 km error submitted their answer within 254 seconds in average (min: 10secs, median: 181 secs, max: 1966secs). The average running time of videos were 52.8seconds for the Ideal 10 set.

Ultimately the purpose of this experiment was the qualifi-

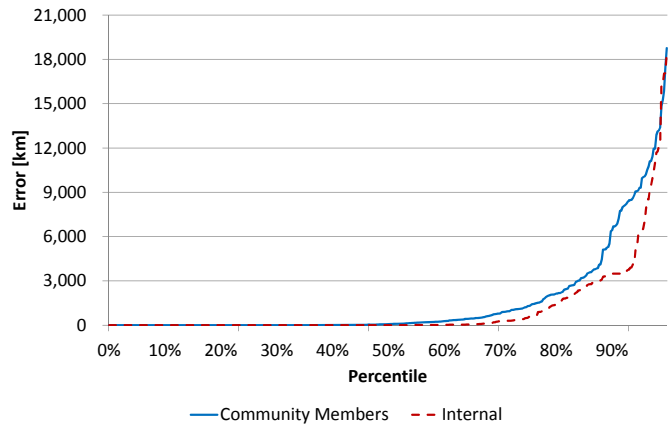


Figure 7: Comparison of the accuracy of highly qualified internal workers (part of the research group) and random qualified (from the community) workers for 40 videos as described in Section 7.

cation of annotators for our ongoing annotation task of un-categorized videos. As we are attempting to produce a baseline for comparison to our automated geolocation systems, we require very high accuracy, thus we set the threshold for qualification at 80 % accuracy. To be scored as getting a correct answer, the video had to be geolocated with 10km of its posted location. This requirement meant that 16 % of participants were qualified to do the actual task. After eliminating the individuals who tried to get the bounty without seriously attempting the task, our acceptance threshold went up to 19 %.

7. COMPARISON TO NON-MECHANICAL TURK RESULTS

Before we conducted the experiments on Mechanical Turk, we built a geotagging web interface² for 30 online workers from the research community which shares similar characteristic with our 'internal workers set'. These invited participants were asked to assign geographical coordinates (latitude and longitude) to 20–30 provided test videos by using textual and audio-visual information. Participants could use external resources such as gazetteers (e.g. GeoNames, Wikipedia) or web mapping service application (e.g. Google Maps) that provides panoramic views from positions along streets. One of the goals of this experiment was to get preliminary feedbacks from the workers and to find bottlenecks in annotating the provided dataset. For this reason it was very interesting that the results of the "professional" annotators sometimes had a better accuracy than the ground truth annotated by the uploader. In the following example, it was easy to find the stadium of the San Francisco Giants on Google Maps by using the tags shown ('dpm563', 'davidspointlessminute', 'sanfranciscogiants', 'battingpractice', 'maysfield', 'southpaw'). Online workers watched the video and tried to see if it was possible to find the precise position of the recording by switching to the aerial view mode. Figure 6 shows some easily recognizable square pillars and a huge stand for detecting the correct angle of the video sequence. The pre-

²<http://geotagging.de.cg/game.php>

cise position of this video sequence was not given by the uploader. The distance between the two points was 39.73 m. For the most precise locations under 1 km the online workers needs an average time of 4 minutes and 48 seconds and the average time is halved for a margin of error of 500 km. All results are shown in table 1.

The distance varies between very small distances of some meters up to 18,000 km because of the sparse representation of geographic information in tags and visual content. As shown in Figure 7, even with the tags provided, the result is worse than the experiment results from the internal volunteers. This is because the videos used in this experiment were picked randomly from the whole dataset. 15% of the videos in the dataset don't have any meaningful metadata and most of these come with home-videos taken in uploader's private properties, thus making the estimation almost impossible.

8. CONCLUSIONS & FUTURE WORKS

Estimating the geo-location of video is not a straightforward "mechanical" task. The task is challenging even for highly-educated, well-travelled and motivated human workers. The task involves collecting and following multimodal clues, i.e. the workers have to utilize one or more of visual, audio, and textual cues from the videos. Results from Section 7 show that only 36.4% and 53% of the answers submitted from research community and the internal volunteers, respectively, had under 10 km error. Our initial approach using Mechanical Turk workers gave us an even lower number with only 24.3% of their answers having under a 10 km error. The goal of this project was to qualify Mechanical Turk workers for the geo-location task. Therefore it was imperative to have the workers' results be similar to those of the motivated volunteers. In order to achieve this, we created an in-depth tutorial, which presented them with rudimentary techniques for approaching this challenging task. After the implementation of the tutorial, our workers achieved 52.8% accuracy, which is almost equal to the performance of the internal volunteer group.

Since we have very large pool of workers on Mechanical Turk, we are able to qualify only those who are able to achieve very high accuracy (19% of our workers had a at least 80% accuracy, which was our threshold for qualification) so that during the task of estimating the location of unknown videos, we can be reasonably certain that they will complete this difficult task with a high degree of accuracy. The bottom line is, it is possible to use crowd sourcing for a very difficult task but one has to be highly careful in the selection of the crowd. In this regard, one creates a pre-selected "elite crowd" much like many systems of higher education do.

We believe that the techniques that we have developed here are applicable to other tasks that require highly skilled crowd-sourced workers. First, it is important to generate a high quality baseline to use for the qualification task. This baseline need not be very large, but it should be created by people who have a good understanding of what the project is. Second, several rounds of internal testing using traditional workers. This will allow you to have feedback on what is confusing about the project in order to develop a tutorial for the crowd workers, and have a baseline of how motivated workers perform. Third, deploy the task to the actual crowd based workers, seek feedback from them, and

compare results to the baseline and motivated workers. Use this information to improve the tutorial, adjust payments if needed, and then repeat the experiment until you have a sufficient number of workers who reach your qualification threshold.

There was some correlation between the amount of time spent on each video and the accuracy. In general, a very brief amount of time spent was an indicator that the worker was not making a real attempt at finding the location, however beyond that brief window there wasn't a strong correlation between time spent and accuracy. For some of the videos, the workers' answers would converge to a point that is more precise than the ground truth given by the uploader. An interesting future experiment would be to try to leverage this to increase the accuracy of the ground truth of the training set. This will be useful for increasing the accuracy of the automatic geo-location estimation system.

Acknowledgments

This research is supported in part by NSF EAGER grant IIS-1128599 and KFAAS Doctoral Study Abroad Fellowship. Human subjects experiments authorized under IRB approval CPHS 2011-06-3325.

We would like to acknowledge Howard Lei for his aid in this work.

9. REFERENCES

- [1] M. Bernstein, E. H. Chi, L. Chilton, B. Hartmann, A. Kittur, and R. C. Miller. Crowdsourcing and human computation: systems, studies and platforms. In *PART 2 ——— Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 53–56, New York, NY, USA, 2011. ACM.
- [2] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17(2):16–21, Dec. 2010.
- [3] D. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 284–291, sept. 2011.
- [4] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.
- [5] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. Jones. Automatic Tagging and Geo-Tagging in Video Collections and Communities. In *ACM International Conference on Multimedia Retrieval (ICMR 2011)*, pages 51:1–51:8, April 2011.
- [6] G. Little and Y.-a. Sun. Human ocr: Insights from a complex human computation process. 2011.
- [7] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008. 10.1007/s11263-007-0090-8.