Zhai, Zhi, Tracy Kijewski-Correa, David Hachen, and Greg Madey. "Haiti earthquake photo tagging: Lessons on crowdsourcing in-depth image classifications." In *Digital Information Management (ICDIM), 2012 Seventh International Conference on*, pp. 357-364. IEEE, 2012

# Haiti Earthquake Photo Tagging: Lessons on Crowdsourcing In-Depth Image Classifications

Zhi Zhai[1], Tracy Kijewski-Correa[2], David Hachen[3], Greg Madey[1]
[1]Department of Computer Science and Engineering
[2]Department of Civil and Environmental Engineering and Earth Sciences
[3]Department of Sociology
University of Notre Dame
Notre Dame, IN-46556, USA

***Abstract*—Facilitated by the latest advances of information technologies, online human computing resources provide researchers unprecedented opportunities to resolve a class of real-world problems that are challenging even to the computer algorithms, and yet manageable to human intelligence if working units are well organized. A problem in this category is image labeling, recognizing and categorizing targets in the images. In this paper, we describe an online platform that leverages human computation resources to resolve an image labeling task – classifying damage patterns in post-disaster photos. The underlying information valuable to us is not only the existence of damage in the image, but also its patterns and severity. We hope this study can provide new perspectives to enhance the design of crowdsourcing projects in future.**

## I. Introduction

Despite the rapid developments of Artificial Intelligence, to this day human intelligence still demonstrates its superiority in a range of areas, such as context retrieval, aesthetic judgment, visual recognition, etc. Meanwhile, along with the progress of information technology, people are increasingly woven into virtual online communities, through which they obtain new knowledge, keep in touch with friends, and/or comment on events occurring in their life circles [11]. This new social phenomenon has motivated us to design innovative web applications that can properly channel scattered human computing power towards solving real-world problems.

In this study, we present a web platform that aims to aggregate human computation resources to tackle a challenging task – classifying damage patterns in post-earthquake photos. In real life, this type of information is needed to manage risks in disaster-prone areas – both in pre-disaster risk reductions and post-disaster damage assessments [13].

Human and economic losses due to large-scale natural disasters are frequently experienced in many populous areas of the world [12][13]. In the aftermath of these disasters, a clear assessment of the damage is desirable for local communities to conduct better damage analysis, infrastructure inspection, remediation and reinforcements.

To this end, we developed a web platform, designed to organize online crowds to collaboratively make efforts towards retrieving structural-damage information from photos collected after the 2010 Haiti Earthquake.

## II. Photo Tagging Platform

In this pilot project, undergraduate students at the University of Notre Dame were recruited as surrogates for citizen engineers. They followed photo tagging procedures developed by researchers from the Department of Sociology and the Department of Civil Engineering, and their online activities were recorded in detail. Over 17 days, the crowd submitted 9318 photo classifications on 400 sample photos.

### A. Procedure Outline

Upon agreeing to a consent form, subjects were directed to a sign-up page, and instructed to create their login credentials. Below, we list 4 major steps in the workflow (interested readers may refer to [15] for detailed procedures).

1) ***Entry Survey*** The purpose of this questionnaire was to collect demographic and attitudinal data from the subjects.
2) ***Introduction Page*** The introduction page describes task background and explains experimental conditions, which was designed to arouse moral sentiments for helping local residents in devastating Haiti Earthquake.
3) ***Tutorials*** Tutorials, as shown in Fig. 1, provide detailed information on how to precisely classify the damage depicted in a photo, and by using hyperlinks, subjects could return to tutorials to reaffirm their understandings about the task.
4) ***Damage Classification*** Subjects received one random photo at a time (a sample photo is seen in Fig. 2), until they completed all of the 400 photos in the database or the allocated 7-day tagging session expired.

### B. Tagging Questions

As shown in Fig. 3, to classify a photo, subjects followed a 5-step damage assessment process. These 5 steps are:

1) ***Image Content*** Determine if an entire structure or only a part of the structure is destroyed in the image.
2) ***Element Visibility*** Identify which elements (*beams*, *columns*, *slabs*, *walls*) of the building are visible and can be assessed.
3) ***Damage Existence*** For each of these visible elements, assess if any of those elements are damaged.

978-1-4673-2430-4/12/$31.00 ©2012 IEEE

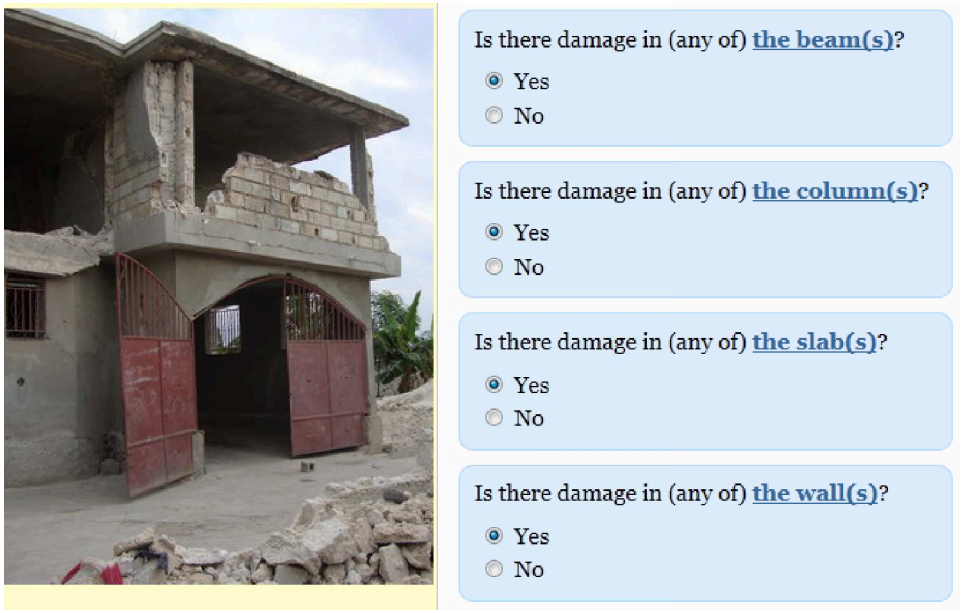


Fig. 2. Web interface of a sample photo.

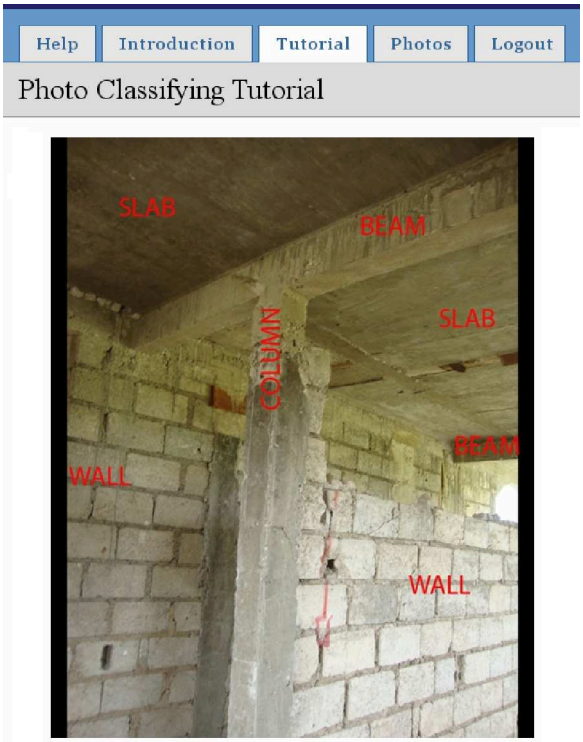


Fig. 1. A sample page of tutorials. Users are required to go through this tutorial before classifying photos, and they can revisit it anytime during the tagging process.

4) ***Damage Pattern*** For each of the elements identified as damaged, indicate the damage pattern.
5) ***Damage Severity*** For each of the elements identified as damaged, appraise the severity of the damage (*Yellow* or *Red*).

These questions are pre-designed by civil engineering professors, and users orderly followed these steps in their tagging practice. However, depending on the damage situations and user perceptions on each photo, for an individual user, tagging process may terminate at an intermediate step if s/he believed that there was no damage on certain building elements.

### C. Defining Ground Truth

3 PhD graduate students in civil engineering (mentioned as *Professionals* hereafter) provided expert judgments on the 400 sample photos. When reviewing Professionals' answers, we realized there were 3 types of consensus:

1) ***Unanimous Consensus*** All 3 Professionals converged to the same answer. Among all of the questions, the unanimous consensus accounts for proximately 30% of answers.
2) ***Majority Consensus*** 2 out of 3 Professionals agreed with each other, and the third Professional diverged from the other two. In the entire question set, the majority consensus accounts for 65%.
3) ***Total Divergence*** 3 Professionals entirely disagreed. 5% of the answers fall into this category.

Note that in real practice of crowdsourcing projects, ground truth is usually not available to data analysts. In other words, ground truth can only be used to evaluate the quality of crowds' work, but not part of the workflow to generate plausible answers.

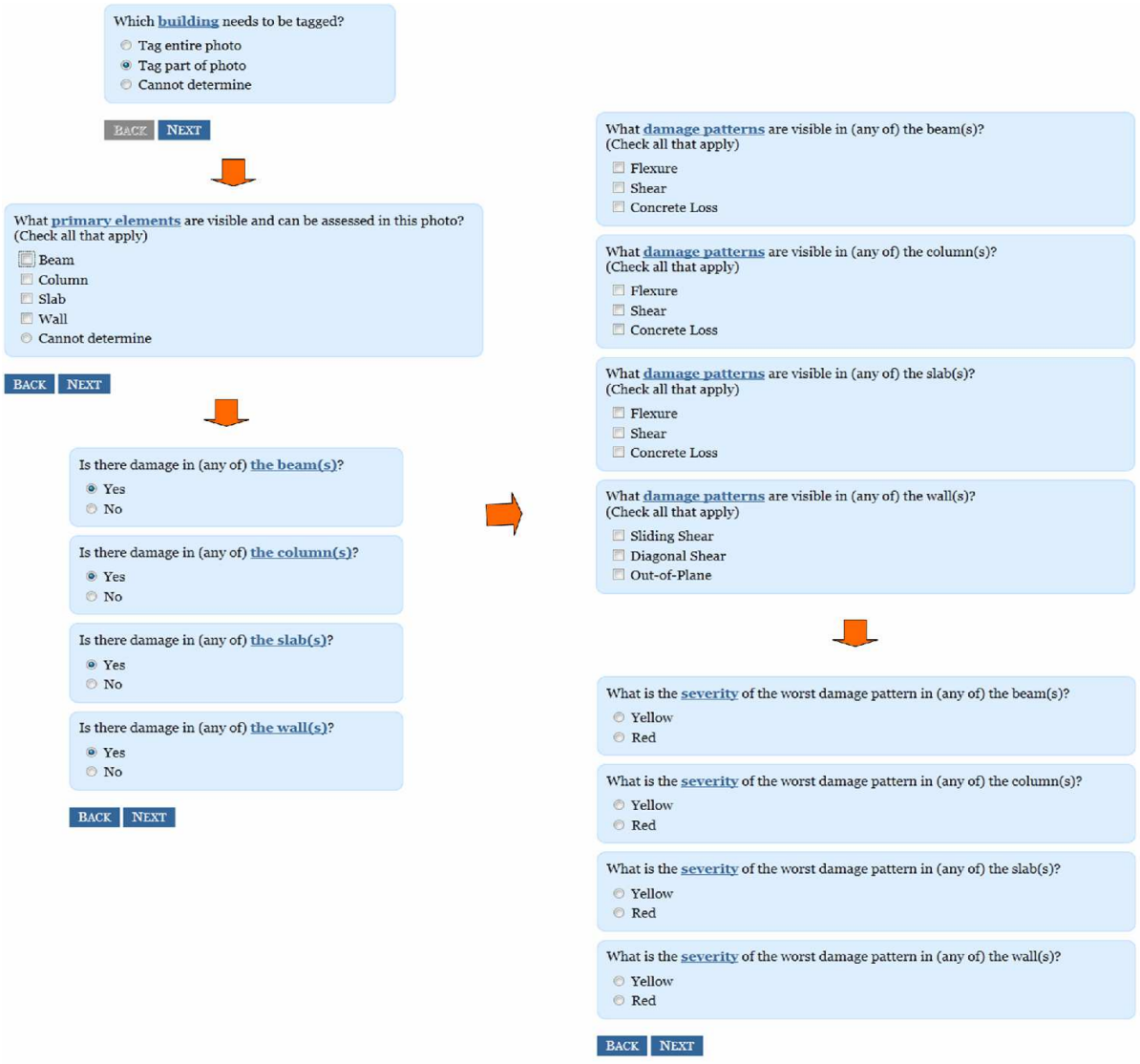


Fig. 3. Workflow. As online photo taggers went deeper along the flow, their answers became increasingly diversified.

## III. Data Cleaning

In [14], the authors discussed anti-cheating strategies in the *Peekaboom* game. In this gaming-with-a-purpose system, if there is a sharp decrease on participants' average playing time, it may indicate that these users start to play the system. In our study, we also observed dramatical decrease on some users' average tagging time, which emphasizes the necessity of data cleansing.

As described in the *Tagging Question* section, the first screen users received has only one high-level question, asking whether the whole structure hit by the earthquake was still recognizable. If a user believed the structure was thoroughly demolished, s/he can simply select the "Cannot Determine" option to proceed to the next photo. In other words, if "Cannot Determine" option was chosen, the tagging process on this photo has terminated at the first step - *Image Content*.

Examining the data, we found a portion of users played the system, considering "Cannot Determine" as a shortcut to explore photos without carefully considering their answers. To an extreme, there were 5 "Cannot Determine" sequences longer than 100 (users consecutively clicked "Cannot Determine" more than 100 times).

To detect these clickers and negate their impact, we experimented with 3 different data cleansing strategies, which are:

1) ***Approach 1: Averaging tagging time*** If an individual average photo tagging time is far below the average tagging time across all users, we suspect this user was not serious, and thus relevant inputs can be discarded.
2) ***Approach 2: "Cannot Determine" proportion*** If the appearances of "Cannot determine" account for more than 75% of the total number of photos that a given user classified, it indicates that this user is careless.
3) ***Approach 3: "Cannot Determine" sequence*** If the length of consecutive appearances of "Cannot Deter-

mine" is longer than 3, this sequence becomes suspicious. It is observed that among the 400 photos, there are a small portion of images, where architectural structures are hardly assessable due to severe damage, to which "Cannot Determine" is a legitimate answer. Yet, it is rare that 3 or more of this type of photos consecutively appear in a row.

In *Approach 1*, we plot the average tagging time across all users in Fig. 4, using equal-width discretization. We can observe that 8 subjects fall into the first bin, which reflects that they spent less time than their peer photo taggers. We can consider these 8 users as mischievous clickers, and remove their inputs from database.

In *Approach 2*, we examine all answers each user has submitted, and subsequently identify a group of clickers according to the percentage of "Cannot Determine"s in their submissions. Then, these clickers' inputs can be easily located and discarded.

The first 2 approaches have the merit of simplicity. The following 2 observations, however, complicate the situation:

1) ***Initial decent work*** For clickers, long "Cannot Determine" sequence did not get started until the clickers found this shortcut. That means the first couple of photos may bear acceptable quality.
2) ***Inferior work after long sessions*** Some serious users suffer low accuracy periods after long tagging sessions, and unintentionally they bring less-trustworthy answers into the system.

Under this circumstance, we expect the third approach has the best performance among the three. In *Approach 3*, the noise is individual tagging sequences, rather than all classifications from a certain group of users.

Table I shows the statistics about the frequency of "Cannot Determine" sequences, and after comparing them with the ground truth, it becomes convincing that these sequences represent error-prone data and should be trimmed. Fig. 5 further illustrates that sequences longer than 3 actually have very low accuracy (below 10%). The third approach, instead of removing entire tagging sets from suspicious users, allows us to keep a good portion of the data that usually occurred at the beginning of users' tagging process, even these users might become careless later on.

Further, we want to use two statistical values to appraise the 3 strategies.

### A. Intraclass Correlation Coefficient (ICC)

Intraclass correlation coefficient (ICC) is a descriptive statistic that measures the similarity between data entries within the group. In our case, since we have categorical data, we set the difference as 0 if two answers are identical, otherwise the difference is 1. ICC equation in our calculation is in its canonical form (in [5], there is more discussion about ICC).

It is observed that both Approach 1 and 3 produce significantly higher intraclass correlation values than initial data set before trimming.

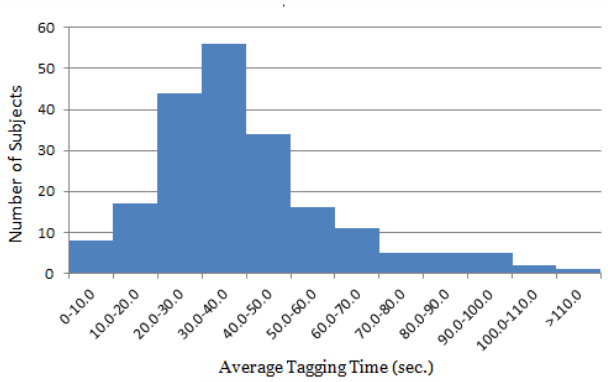


Fig. 4. Subject distribution on average tagging time. Equal-width discretization.

ICC is a proper indicator of crowd homogeneity. Also, we want to measure the crowd consensus' accuracy, which is achieved by the following metric - crowd consensus score.

### B. Crowd Consensus Score

We collected opinions from 3 Professionals. According to this ground truth, the maximal points crowd can collect across all 400 photos is 4905. We then can compare the answers from the crowd and the answers from the ground truth: for each question, if the crowd consensus and Professionals' answer are the same, the crowd earns one point. Otherwise, they do not receive any point on this question.

In more complex scenarios, if there is a 2-way tie in the crowd consensus, crowds will receive a half point, and if there is a 3-way tie, the crowds will receive one third point.

In this measure for accuracy, the crowd received 2750 without trimming, 3245 after Approach 1 running through, and 3487 after Approach 3 applied alone. Approach 2 generates inferior results to the data set without pruning, and the combination of Approach 1 and 3 did not produce more desirable results.

## IV. Bias Resource I - Professionals

When calculating the crowd consensus, we implemented equal-weight voting - one vote represents one person, and all votes have equal weights. When we tested crowd consensus against the ground truth - the unanimous and majority consensus from Professionals, the crowd had 71% accuracy, which was barely satisfactory.

In the post-experiment interview with Professionals, they indicated that, when reviewing the photos, oftentimes they tended to exert their expertise to evaluate the damage patterns behind the scenes.

Fig. 6 illustrates an example: when classifying a given photo, Professionals may have different emphases: either being comprehensive or conservative. In contrast to traditional photo classifying projects, where goals are usually to judge the existence of certain targets and human biases can be effectively rectified by providing detailed tutorials and instructions, in

TABLE I
STATISTICS OF "CANNOT DETERMINE" SEQUENCES

| length | Appearance | Frequency (%) | Accuracy (%) |
|---|---|---|---|
| 1 | 541 | 72.72 | 78 |
| 2 | 103 | 13.84 | 50 |
| 3 | 37 | 4.97 | 9 |
| 4 | 21 | 2.82 | 6 |
| 5 | 9 | 1.21 | 7 |
| 6 | 3 | 0.40 | 7 |
| 7 | 1 | 0.13 | 4 |
| 8 | 5 | 0.67 | 8 |
| 9 | 3 | 0.40 | 8 |
| 10 | 1 | 0.13 | 8 |
| 10∼20 | 6 | 0.80 | 3 |
| 20∼30 | 4 | 0.54 | 6 |
| 30∼40 | 2 | 0.27 | 3 |
| 40∼50 | 0 | 0 | 0 |
| 50∼60 | 1 | 0.13 | 4 |
| 60∼70 | 0 | 0 | 0 |
| 70∼80 | 0 | 0 | 0 |
| 80∼90 | 0 | 0 | 0 |
| 90∼100 | 1 | 0.13 | 3 |
| >100 | 6 | 0.80 | 1 |
| Total | 744 | 1 | 51.99 |

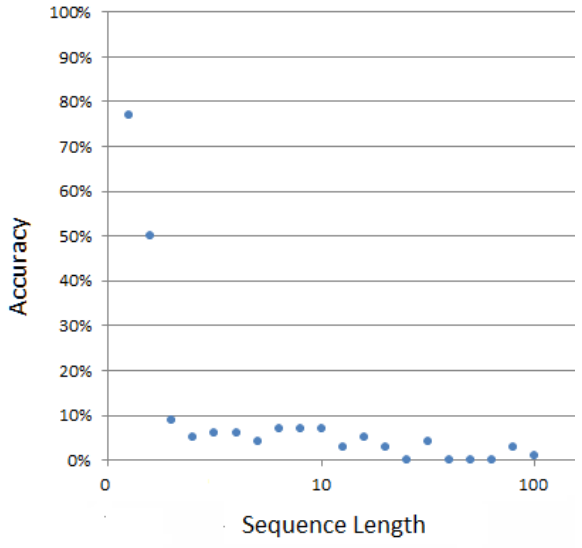


Fig. 5. Accuracy vs. "Cannot Determine" Sequence Length. We can observe the rapid decrease in accuracy as sequences grow. The data points are from the *Accuracy* column in Table I.


Fig. 6. While 3 professionals achieved agreements on the *Column* damage in Area 1 and *Wall* damage in Area 2, they took different positions on whether there was a *Beam* damage in Area 3. (Professionals were not required to draw frames while tagging photos)

post-earthquake photos misjudgments can go opposite directions, as damages demonstrated themselves in volatile situations. As a result, various types of flaws/damage are unlikely to be fully addressed in tutorials, and thus relevant biases usually cannot get satisfactorily neutralized.

This concern is evidenced by the following statements from the professionals:

***Comprehensive***:

- *Could not fully see what happened to the walls, but I know the damage exists.*
- *Difficult to decide on the damage pattern: shear vs. flexure, so I chose both 2 damage types out of three.*

***Conservative***:

- *Pretty much everything is damaged, but hard to tell what is what though. So, I selected Beam, Slab, and Wall that can be clearly seen.*
- *Again, all are damaged, but it's hard to differentiate building parts from the photo. I decided to leave Column out.*

These responses imply that some professionals were rigorous on providing defensible answers, while others strove to cover more details. When we took into account the possible biases from Professionals' judgments and recalculated the crowd performance, the crowd's accuracy increased to 84%, significantly higher than was the case before redefining the ground truth.

## V. BIAS RESOURCE II - OVER CLASSIFICATION

In the 2-week period, 132 out of 204 users have classified at least 10 photos. Because of this willingness, it was easier for project organizers to gather sufficient amount of human resources to conduct the research. On the other hand, users had the tendency to be excessively careful covering convoluted damage traces. This trend is reflected on 3 observations:

1) ***Building Structure*** Besides the prominently damaged part(s), building parts with non-essential flaws are also pointed out as damaged.
2) ***Damage Pattern*** Vague or trivial damage patterns are considered as being substantial, no difference from the prominent and major patterns.
3) ***Condition Severity*** Over-estimated severity of the damage when the structure is still reparable.

These crowds biases are further examined against the ground truth, and discrepancies mainly demonstrate in 3 aspects.

1) ***Unclear Building Part(s)*** The crowd believed there was damage existing on a building part, but ground truth indicated no damage on this element. For example, in Fig. 7 the crowd had consensus on Column damage, resonating with Professionals. However, the crowd also classified the wall attached to the column as a damaged building part, which differed from Professionals' opinion. In this case, we regard this question as an instance of over classifying.
2) ***Unclear Damage Pattern(s)*** The crowd believed there were more than one damage patterns occurring, but ground truth indicated only a subset of them exist.
3) ***Over-Estimated Severity*** The crowd believed the structure was destroyed (marked as *Red*), but ground truth indicated although it was affected, it was still reparable and useful (marked as *Yellow*).

Note that, contrary to the bias resources in section IV, the above 3 categories of bias lean towards the same direction: over-classifying as opposed to under-classifying, and hence could be negated by more informative tutorials and instructions.

After taking care of the over-classification issue, the crowd further increases its performance by 9%, reaching 91.6% accuracy. In this section, our goal is to carve out the most trustworthy results we can depend on.

## VI. Lessons and Experiences

The questions we asked about each photo are programmed, but the cohort of users answer these multiple-option questions according to their perceptions about the photo. This challenges our quality control strategies. In future design, there are 3 techniques we would consider to take so as to improve classification quality:

### A. Blending objective questions with subjective questions.

By inserting objective questions into the questionnaire, such as "what is the magnitude and epicenter of the earthquake" or "where is the most populous area in the country", we can trace if users have acquired basic knowledge about the task. Also, these objective questions with firmly verifiable answers will make it clear to users that their answers can and will be assessed in data analysis phase - preventing gaming behaviors, potentially increasing effort [6], and helping the project organizer preclude inferior inputs.

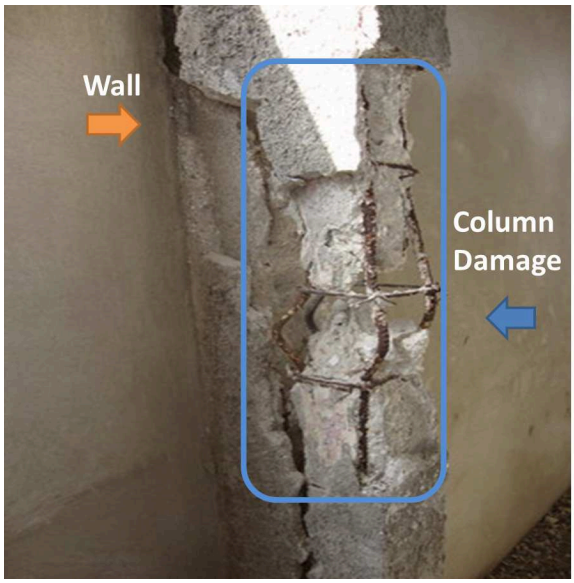


Fig. 7. Besides the salient *Column* damage, the crowd also included the wall attached to the column as a damaged building part. (Participants were not required to draw frames while tagging photos)

### B. Measuring the level how asserted they are about answers.

As explained in the last section, in experiments where moral responsibility is the major motivation, users are usually inclined to excessively exert their enthusiasm, which is reflected as over-classification in our case. In future design of social-benefit projects, e.g. risk reduction, environment surveillance, etc., confidence levels of the users about their opinions should be taken into consideration. Users are expected to submit their answers as well as how sure they are about their answers. This way, designers can make pruning-retaining decisions in accordance to hierarchical confidence levels.

### C. Providing users morality encouragement.

In this study, users participated not for monetary rewards but moral fulfillment, and thus there are several approaches we can take advantage of to motivate users with stronger encouragement.

- Send them thank-you notes on behalf of the local residents suffering natural disasters.
- Acknowledge taggers' efforts, and feature their contributions in social media, such as the school newspaper or websites.
- Accolade users with token/kudos recognitions along the working process, such as virtue medals and stars.

## VII. Related Work

In previous sections, we addressed some concerns and considerations in this study, which aims to harness the collective intelligence for disaster risk reduction. In this section, we explain a list of previous work in the literature that we drew inspirations from.

### A. CrowdFlower and Samasource

Thanks to the development of the Internet, compared to the traditional way in disaster relief, online resources are more accessible and pervasive. There have been successful practices that leveraged efforts of virtual communities to provide urgent service to devastated areas.

Right after the Haiti Earthquake, crowdsourcing platform Crowdflower and Samasource coordinated online crowds to offer SMS message real-time translations that effectively overcame the language barriers between the aid workers awaiting information to dispatch personnel and the disaster areas needing help [3].

Compared to these early disaster relief efforts, our pilot project attempts to support local communities from another perspective - post-disaster assessment and risk reduction.

### B. Galaxy Zoo

Releasing a mass amount of photos online and appealing crowds to contribute time and expertise is a practical approach to classify images. In Galaxy Zoo [8] project, astronomers from Oxford University established a website providing (1) Astronomical photos collected from telescopes, (2) tutorials for users without professional training to obtain basic knowledge, and (3) interface through which users can submit their answers.

Like Galaxy Zoo, when developing the photo-tagging website, we especially emphasized the tutorials to be well organized and structured for beginners to follow. Besides, the 5-layer questions of each photo are intended to retrieve comprehensive information out of each photo-user pair, which presents a great potential to generate otherwise unrevealed knowledge because of its depth, compared to the similar image classification work conducted in [1].

### C. ZoneTag

ZoneTag is a photo labeling project that heavily relies on geo-information supported by users' smart phones [9]. Every time a user in the field takes a picture, in real time s/he automatically gets a list of tag suggestions from the central server connected to *Flickr*. Therefore, presented by the labels suggested based on historical records, a user has candidate labels to name their new photos instantly.

In ZoneTag, suggestions are based on crowd history, which may lose accuracy in the new situation where disasters perhaps have defaced the area. However, it nonetheless highlights an innovative approach - attaching meta-data or initial assessment to photos right at the venue, providing first-hand information for later processing.

## VIII. Conclusions and Future Work

In this paper, we introduced a pilot project – Haiti Earthquake Photo Tagging – where online volunteers fulfill basic human computations [7]. Using statistical tools, we try to shed new light on analyzing data and rectifying biases commonly observed in crowdsourcing projects. By presenting lessons learned and experiences obtained from the experiment, we provide insight and guidelines for future citizen engineering project designs.

For projects that strive to tap into online unidentified crowds, quality control is always pivotal to achieve trustworthy results. In this project, we use crowd consensus self-check and statistical pruning to achieve high trustworthiness. Other strategies worth further investigations include *Ground Truth Seeding* [10], *Multilevel Review* [2] and *Defensive Task Design* [4].

How to effectively recruit and motivate crowds is another related research topic. In this study, during the recruiting phase we did not encounter particular difficulty in enlisting college students to participate. In future research, however, to scale up this crowdsourcing system beyond the college campus, we need to explore different motivating mechanisms such as: *entertainment*, *camaraderie encouragement* [6], *social recognition*, *intrinsic satisfaction*, and possibly the combination of the above.

Regarding the user recruitment, an issue that may raise is the representativeness of experiment subjects. College students are generally believed to be individuals with higher education level and stronger moral motivations, which hardly are the common characteristics of online workforce. To address this concern, we would like to extend our research to commodity crowdsourcing platforms such as Amazon Mechanical Turk (AMT), and we believe a comprehensive comparison between experimental data collected from these two different platforms - AMT and our campus platform - would bring more insight and perspectives to citizen engineering research community.

## Acknowledgment

The research presented in this paper was supported in part by an award from the National Science Foundation, under Grant No. CBET-0941565 for a project entitled *Open Sourcing the Design of Civil Infrastructure (OSD-CI)*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank Jenny Vaydich and Zack Kertcher for valuable contributions, and the paper reviewers' constructive comments and suggestions.

## References

[1] ImageCat. http://www.imagecatinc.com/, Retrieved Aug. 2011.

[2] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 313–322, New York, New York, USA, 2010. ACM.

[3] L. Biewald. Massive multiplayer human computation for fun, money, and survival. *XRDS*, 17(2):10–15, Dec. 2010.

[4] C. Callison-Burch and M. Dredze. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 1–12, Los Angeles, California, 2010.

[5] R. A. Fisher. Statistical methods for research workers. *The Eugenics Review*, 18(2):148–150, 1926.

[6] A. Kittur. Crowdsourcing, collaboration and creativity. *XRDS*, 17(2):22–26, Dec. 2010.

[7] E. Law and L. von Ahn. *Human Computation (Synthesis Lectures on Artificial Intelligence and Machine Learning)*. Morgan & Claypool Publishers, 1 edition, July 2011.

[8] C. J. Lintott1, K. Schawinski1, A. Slosar1, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389:1179–1189, 2008.

[9] M. Naaman and R. Nair. Zonetag's collaborative tag suggestions: What is this person doing in my phone? *IEEE MultiMedia*, 15(3):34–40, July 2008.

[10] A. Quinn and B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1403–1412, Vancouver, BC, Canada, 2011.

[11] C.-C. Shih, T.-C. Peng, and W.-S. Lai. Mining the blogosphere to generate local cuisine hotspots for mobile map service. In *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, pages 1 –8, nov. 2009.

[12] United Nations International Strategy for Disaster Reduction (UNISDR). 2009 global assessment report on disaster risk reduction - risk and poverty in a changing climate. 2009.

[13] United Nations International Strategy for Disaster Reduction (UNISDR). 2011 global assessment report on disaster risk reduction 2011 -revealing risk, redefining development. 2011.

[14] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, Montreal, Apr. 2006. ACM Press.

[15] Z. Zhai, D. Hachen, T. Kijewski-Correa, F. Shen, and G. Madey. Citizen engineering: Methods for "crowd sourcing" highly trustworthy results. In *Proceedings of the Forty-fifth Hawaii International Conference on System Science (HICSS-45)*, Maui, HI, USA, Jan. 4-7 2012.