Published in final edited form as:

Curr Opin Ophthalmol. 2016 May ; 27(3): 256–261. doi:10.1097/ICU.0000000000000251.

Crowdsourcing: An overview and applications to ophthalmology

Xueyang Wang, BS, Lucy Mudie, MBBS MPH, and Christopher J. Brady, MD

Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland

Abstract

Purpose of review—Crowdsourcing involves the use of the collective intelligence of online communities to produce solutions and outcomes for defined objectives. The use of crowdsourcing is growing in many scientific areas. Crowdsourcing in ophthalmology has been used in basic science and clinical research, however it also shows promise as a method with wide-ranging applications. This review presents current findings on the use of crowdsourcing in ophthalmology and potential applications in the future.

Recent findings—Crowdsourcing has been used to distinguish normal retinal images from images with diabetic retinopathy; the collective intelligence of the crowd was able to correctly classify 81% of 230 images (19 unique) for USD\$1.10 per eye in 20 minutes. Crowdsourcing has also been used to distinguish normal optic discs from abnormal ones with reasonable sensitivity (83–88%), but low specificity (35–43%). Another study used crowdsourcing for quick and reliable manual segmentation of OCT images. Outside of ophthalmology, crowdsourcing has been used for text and image interpretation, language translation, and data analysis.

Summary—Crowdsourcing has the potential for rapid and economical data processing. Among other applications, it could be used in research settings to provide the “ground-truth” data, and in the clinical settings to relieve the burden of image processing on experts.

Keywords

crowdsourcing; ophthalmology; Amazon Mechanical Turk

Introduction

Crowdsourcing is generally described in academic literature as “an online, distributed, problem-solving, and production model that uses the collective intelligence of networked communities for specific purposes”(1). The corresponding author, CJB, has previously described the application of crowdsourcing for prompt interpretation of retinal images (2). In that study, the Amazon Mechanical Turk (AMT) platform was used to crowdsource grading of fundus photos for diabetic retinopathy; for USD\$1.10 per eye, workers were able to correctly grade images as normal or abnormal 81.3% of the time, workers spent on average 25 seconds on each image (2) (Figure 1). Crowdsourcing offers a novel method for data

Corresponding author: Christopher J. Brady, Assistant Professor of Ophthalmology, Wilmer Eye Institute, 600 N. Wolfe St, Maumenee 711, Baltimore, MD 21281, 410-502-2789, cbrady5@jhmi.edu.

Conflicts of interest: none

processing and could be applied to many areas in ophthalmology. Herein, we review the recent literature regarding crowdsourcing and its use in biomedical research more broadly and vision research specifically.

Distributed Human Intelligence

Distributed human intelligence is a subset of crowdsourcing that involves the division of a large task into its smaller components, which are then completed by individuals, so that only collectively is the entire task completed (3). AMT is an online marketplace for distributed human intelligence tasks; a task administrator (“Requestor”) uploads several small, discrete tasks, known as human intelligence tasks (HITs), which become available for completion by thousands of potential workers (“Turkers”). Common HITs involve translating words or labelling images. However, HITs may involve anything from searching the internet to find the correct website address of a company, or watching a video to rate the level of sarcasm in the dialogue; often these tasks remain challenging or inefficient for artificial intelligence algorithms. Most HITs can be completed in seconds to minutes, and once the Turker completes their HIT, it is sent to the Requestor for approval. Once approved, the Turker is generally paid \$0.01–\$0.25 for completing the HIT, depending on the complexity of the task. Typically, a Requestor will upload a group of similar HITs, known as a “batch”, and depending on the type and complexity of the batch, as well as the payment offered by the Requestor, the batch may be completed within minutes to hours at a very economical cost.

To sustain the integrity of the marketplace, AMT is a reputation-based economy; the most sought-after HITs are only available to Turkers with a high approval rating (workers who have a high proportion of their HITs approved). Likewise, new Requestor’s HITs may be avoided by Turkers with high ratings until the Requestor develops a reputation for fairness in approving and rejecting work. HITs may also have a qualification test associated with them, the qualification may be location or skills based, or even a stipulation that the Turker has *not* completed the particular HIT before (i.e., is naïve to the task).

Turkers work anonymously, however demographic studies of Turkers have been performed, and there is now a web-based tool that tracks, in real time, the demographics of the population of Turkers online in the AMT marketplace (4). One study found 46.8% of Turkers were located in the U.S., 34% were in India, and the remaining 19.2% in 64 other countries (5). In the U.S., the majority of Turkers were women who used AMT as source of supplemental income, while in India, Turkers were more likely to be men earning their primary income (5). However, the demographics of Turkers completing HITs varies with the day and time; for example, at 5pm U.S. Eastern time, Turkers are more likely to be located in the U.S., while at 3am U.S. Eastern time, they are more likely to be located in India (4). In both India and the U.S., Turkers are generally born in the 1980s, and the majority hold at least a Bachelor’s degree (5). Most workers completed 20–50 HITs per week for USD\$1–5 (5). Among U.S. Turkers, monetary reward remains the primary impetus for completing HITs, although the motivations may be more complex; many of the surveyed Turkers included entertainment and education as reasons for being a worker (5).

AMT is a dynamic and complex ecosystem of distributed human intelligence, albeit a powerful one. AMT has brought crowdsourcing to the fingertips of the masses; it has instantaneously and conveniently connected those with a task to be accomplished to the workers who are willing to complete it. The capabilities of AMT has many applications for biomedical research and beyond.

Crowdsourcing as Tool for Biomedical Research

Crowdsourcing has been used in fields of scientific research ranging from basic science to public health. The crowd has interpreted and annotated medical images and documents to contribute to databases for future research (6–8). The crowd has also assessed the skills of surgeons and played games to map retinal ganglion cell neurons in mice (9–13). In almost all studies, the cost and time of using the crowd were much less than that of using experts. Such reports portray crowdsourcing as a tempting platform to process large batches of information while saving valuable resources.

Text Interpretation

Crowdsourcing has been used for behavioral research, with Turkers from AMT acting as substitutes for expert judges in audio transcription, document comparison, and language processing (14). Azzam and Harman showed Turkers were able to consistently rate the most important point in a transcript and identify supporting text segments from the transcript to explain their ratings (15). In similar study, Turkers highlighted words and phrases to indicate disease processes in PubMed abstracts after passing a qualification test and completing annotation trainings (6). In a span of 9 days and for a cost of under \$640, Turkers annotated 589 abstracts 15 times and produced a disease mention annotation corpus similar to the gold standard (NCBI Disease corpus) (6).

Image interpretation

The developers of ImageCLEFmed, an automatic, medical image modality classification program, used Crowdfunder, another crowdsourcing platform, to improve and expand a training set of images, so that the program could learn to correctly classify a diverse range of images (7). AMT has also been used to produce reference images to verify new algorithms in computer-assisted minimally invasive surgery (8). In under 24 hours, at a cost of \$100, 10,000 annotations were obtained from Turker grading, which were deemed expert-quality after clustering.

Games

The crowd have also played interactive computer games to achieve research objectives, with successful projects including FoldIt for protein folding and MalariaSpot for image analysis (9, 10). Recently, Waldispuhl et al. developed a human-computing game called Ribo to improve the accuracy of sequence alignment for noncoding RNA, which is usually done manually in a time consuming fashion by highly trained individuals (11). The nucleotides were represented with colored blocks that could be moved within a grid, and the players were asked to align the like-colored blocks within the same columns to reveal similar sequences. The average Expect (E)-values of the alignments from the Ribo game were less

than the average E-values from an established sequence alignment database, indicating greater precision by the crowd. This proof-of-concept study showed that crowdsourcing can be harnessed to improve the accuracy of public RNA alignment databases.

Another game in the basic science crowdsourcing world is EyeWire (12). Investigators converted the mapping of individual retinal ganglion cells into a web-based game, which engaged the crowd by publishing users' high scores and allowing users to vote on "consensus" segmentation. The players of EyeWire were committed to the game, with some logging thousands of neuron "cubes." Their accuracy also improved over hours of practice, yielding a three-dimensional map of the branching patterns of mouse retinal neurons in a relatively short amount of time. This research demonstrated that the crowd is not only a powerful tool for scientific discovery but also a labor force motivated as much by fun as by financial gains.

Assessment

Crowdsourcing has been used to evaluate surgical skills, yielding results similar to evaluations performed by expert surgeons. White et al. generated videos of surgeons performing dry-laboratory robotic surgical skill assessment tasks, which were then graded by 3 experts as well as by Turkers using a standardized grading tool (13). The crowdsourced scores had a high degree of agreement with expert scores, with the crowd costing \$16.50 to grade each video and the experts costing \$54–\$108 for the same task. The crowd was biased toward being more critical of the top performing surgeons than the experts were. Turkers have also been asked to assess surgical skills in videos of live, porcine robot-assisted urinary bladder closures (16). In this study, Turkers agreed with expert scoring ($\alpha > 0.90$) and completed the task in less than 5 hours, more than 13 days faster than the surgeon graders (16). In a separate study, Turkers consistently generated the same rank order of lower scoring surgeons as the expert graders in reviewing videos of robotic-assisted radical prostatectomy (17). Once again, Turkers completed the task far quicker than experts; 2,531 video ratings were made by Turkers in 21 hours, while the experts took 15 days to complete 318 ratings.

The applications of crowdsourcing in medical research are many. To date, most studies have piloted the use of the crowd to complete tasks usually done by experts, and nearly all have been proof-of-concept reports or first expansions from the baseline. The same phenomenon is occurring in ophthalmology, where researchers are beginning to adapt crowdsourcing as a tool for data processing and analysis.

Crowdsourcing in Ophthalmology

In ophthalmic research, image analysis is often a time consuming process performed by trained professionals. Crowdsourcing is being explored as a way to lessen the burden on experts and process large batches of images. Mitry et al. used AMT to grade 100 fundus photographs with a wide variety of findings as normal or abnormal (18). The Turkers were trained by reading background information on the nature of the image and descriptions of abnormal features found in the images. They could also consult 2 labeled examples of normal fundus photos. The researchers ran 4 different study designs, each time varying the

amount of prior experience a Turker needed to have and the compensation per HIT. In under 24 hours, they completed 2,000 classifications for a total cost of \$60. They found that Turkers were able to identify severely abnormal photos with a sensitivity of >98% and a high AUC (range 0.819–0.915). However, the sensitivity was lowered to between 61–72% when Turkers were asked to distinguish normal from mildly abnormal images. Having more stringent requirements for Turker qualification and raising compensation did not increase the sensitivity of the grading.

Next, Mitry et al. sought to determine whether Turkers could identify images with glaucomatous changes (19) (Figure 2). Turkers graded 127 images of optic discs as normal or abnormal, with each image being graded by 20 different Turkers. The first batch had no restrictions on Turker past experience, while the second only allowed Turkers who had completed more than 500 HITs and had >90% approval rating to complete the tasks. Both schemes were run twice, yielding similar findings. Sensitivity ranged between 83–88%, but specificity was low, ranging from 35–43%. The AUC ranged from 0.62–0.64, which was lower than expert grading (0.86) and automated grading (0.88). Turkers were more confident in their classifications of abnormals, with 100% of the glaucomatous retinal images graded correctly while only 8–36% of the normal images were accurately classified. While crowdsourcing yielded efficient ophthalmic image grading, with 2,540 unique classifications of 127 images completed in several hours at “minimal cost,” it is not yet comparable to expert grading and has a high rate of false positives.

Diabetic retinopathy fundus photos have also been graded by Turkers with 2-category (normal vs. abnormal), 3-category (normal, mild/moderate, and severe), and 4-category (normal, mild, moderate, and severe) grading (2). The 2-category system yielded a sensitivity of 100% and specificity of 71.55%, with 81.3% of the images being classified correctly. However, the accuracy decreased to 64.5% and 50.9% with 3-category and 4-category grading, respectively. Unlike Mitry, et al, this group found that accuracy improved after more strict Turker selection and further training and concluded Turkers could grade fundus photos of diabetic patients as normal or abnormal with reasonable accuracy and high speed.

In addition to image analysis, Turkers have been employed for OCT segmentation (20). Macular OCT images were distributed through AMT, and Turkers were paid \$0.01 to delineate the inner limiting membrane (ILM)-retina interface and the retinal pigment epithelium (RPE) line after consulting example images. This was done quickly, with the average time to completion being 30.83 seconds. More than 9,200 data points were collected, and the total cost of segmentation per OCT image was \$0.18. The Turkers had a high degree of agreement with each other, Pearson’s correlation of inter-rater reliability was 0.995 ($p<0.0001$). Agreement with expert or automated segmentation were not reported. With advanced web coding, AMT is a robust platform for manual segmentation of OCT images.

Future Applications of Crowdsourcing

Leveraging the collective intelligence of the crowd has numerous and diverse potential applications. In ophthalmology, crowdsourcing may enhance research capabilities by providing a convenient and economical way to collect, integrate and analyze data. For example, crowdsourcing could be used to provide the “ground truth” on data which could then be used to test and improve other novel methods of data management, such as artificial intelligence algorithms. In the clinical practice of ophthalmology, crowdsourcing could be used as aid in screening large populations for ophthalmic disease. Rapid, inexpensive and accurate interpretation of images through crowdsourcing could reduce the number of images needing specialist grading, and complement screening and telehealth programs.

Conclusion

Crowdsourcing is a novel and convenient approach for data processing. Recent studies have demonstrated that crowdsourcing may be used as a rapid and accurate method to distinguish normal retinal images from images with diabetic retinopathy and glaucoma. The AMT crowdsourcing marketplace has also been used to promptly and reliably complete manual segmentation on OCT images. In other areas of research, crowdsourcing has proven useful for text and image interpretation, and data analysis. Future research should be performed to evaluate ways of maximizing accuracy of data processing by crowdsourcing, as well as to expand the application of harnessing collective intelligence.

Acknowledgments

none

Financial support and sponsorship: Dr. Brady was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number KL2TR001077.

References

1. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. *Am J Prev Med.* 2014; 46:179–187. [PubMed: 24439353]
2. Brady CJ, Villanti AC, Pearson JL, et al. Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. *J Med Internet Res.* 2014; 16(10):e233. [PubMed: 25356929]
This is the most recent article describing crowdsourcing for grading of images with diabetic retinopathy. This paper demonstrated that AMT could be used for rapid and economical grading of fundus photos for diabetic retinopathy with excellent sensitivity and reasonable specificity.
3. Brabham, DC. *Crowdsourcing*. MA: MIT Press; 2013.
4. Ipeirotis, PG. [Accessed 11-6-2015] mturk tracker. at <http://www.mturk-tracker.com/#/general>
5. Ipeirotis, PG. *Demographics of mechanical turk*. (CeDER Working Paper-10-01). New York University; 2010.
6. Good BM, Nanis M, Wu C, Su A. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac Symp Biocomput.* 2015:282–293. [PubMed: 25592589]
7. Garcia Seco de Herrera A, Foncubierta-Rodríguez A, Markonis D, Schae R, Muller H. Crowdsourcing for medical image classification. *Swiss Medical Informatics.* 2014:30.
8. Maier-Hein L, Kondermann D, Ross T, et al. Crowdtruth validation: a new paradigm for validating algorithms that rely on images correspondences. *Int J CARS.* 2015; 10:1201–1212.
9. Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. *Nature.* 2015; 466(7307):756–760. [PubMed: 20686574]

10. Luengo-Oroz MA, Arranz A, Frean J. Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *J Med Internet Res*. 2012; 14(6):e167. [PubMed: 23196001]
11. Waldispühl J, Kam A, Gardner P. Crowdsourcing RNA structural alignments with an online computer game. *Pac Symp Biocomput*. 2015:330–341. [PubMed: 25592593]
12. Kim JS, Greene MJ, Zlateski A, et al. Eyewirers. Space-time wiring specificity supports direction selectivity in the retina. *Nature*. 2014; 509:331–336. [PubMed: 24805243]
13. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS. Crowd-Sourced Assessment of Technical Skill: A Valid Method for Discriminating Basic Robotic Surgery Skills. *J Endourol*. 2015; 29(11):1295–1301. [PubMed: 26057232]
14. Mason W, Suri S. Conducting behavioural research on Amazon's Mechanical Turk. *Behav Res Methods*. 2012; 44(1):1–23. [PubMed: 21717266]
15. Azzam T, Harman E. Crowdsourcing for quantifying transcripts: An exploratory study. *Eval Prog Plan*. 2015; 54:63–73.
16. Holst D, Kowalewski TM, White LW, et al. Crowd-Sourced Assessment of Technical Skills: Differentiating Animate Surgical Skills Through the Wisdom of Crowds. *J Endourol*. 2015; 29(10):1183–1188. [PubMed: 25867006]
17. Peabody J, Miller D, Lane B, et al. Wisdom of the crowds: Use of crowdsourcing to assess surgical skill of robot-assisted radical prostatectomy in a statewide surgical collaborative. *J Urol*. 2015; 193(4S):e655–e656.
18. Mitry D, Peto T, Hayat S, et al. Crowdsourcing as a Novel Technique for Retinal Fundus Photography Classification: Analysis of Images in the EPIC Norfolk Cohort on Behalf of the UKBiobank Eye and Vision Consortium. *PLoS One*. 2013; 8(8):e71154. [PubMed: 23990935]
19. Mitry D, Peto T, Hayat S, et al. Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. *PLoS One*. 2015; 10(2):e0117401. [PubMed: 25692287] This is the first article to describe the use of crowdsourcing for grading of glaucomatous features of digital fundus photos. The authors showed that AMT could be used to distinguish normal optic discs from glaucomatous ones. This study reported highly efficient optic disc grading by Turks, with good sensitivity but low specificity.
20. Lee AY, Tufail A. Mechanical Turk based system for macular OCT segmentation. *IOVS*. 2014 Meeting Abstract.

Key points

- Crowdsourcing has the potential to reliably process large amounts of data in a very rapid and inexpensive way.
- Crowdsourcing has been used to distinguish normal retinal images from images with diabetic retinopathy and glaucoma.
- Crowdsourcing has also been used to reliably complete manual segmentation on OCT images.
- Crowdsourcing has the potential to provide the “ground truth” on data which could then be used to test and improve other novel methods of data processing, such as artificial intelligence algorithms.
- Rapid, inexpensive and accurate interpretation of images through crowdsourcing could reduce the number of images needing specialist grading, and complement screening and telehealth programs.

Figure 1. Screen capture of the authors’ Amazon Mechanical Turk (AMT) interface used for grading of fundus images for abnormalities associated with diabetic retinopathy.

Eye Photo Grading Tutorial and Quiz Requester: Christopher Brady Qualifications Required: Masters has been granted		Reward: \$0.05 per HIT Duration: 60 minutes
--	--	--

We will be posting a series of HITs grading eye photos for glaucoma. Some will require you complete this training/qualification. You will need to view the video tutorial and then grade the 5 test images.
 This task is difficult, but all good faith efforts at this HIT itself will be accepted.
 Users who pass the quiz (Question 1 only, 4 out of 5 correct) will get the first qualification and a \$3 bonus.

Follow this link to view the training video: [Glaucoma Grading Tutorial](#)

After viewing, please grade the 5 images below. You may refer to the video as much as you like.

<div style="margin-top: 5px;"> <input type="button" value="Click for magnified image"/> </div>	<p>1) Label this eye:</p> <p> ☐ Unlikely glaucoma ☐ Possibly glaucoma ☐ Probably glaucoma ☐ Definite glaucoma </p> <p>*) Please rate the quality of the image PLEASE ANSWER, NOT GRADED FOR QUIZ</p> <p> ☐ Poor ☐ Acceptable ☐ Excellent </p>	<p>PLEASE ANSWER, NOT GRADED FOR QUIZ</p> <p>*) What is the size of the nerve?</p> <p> ☐ Small ☐ Medium ☐ Large </p> <p>*) What is the cup-to-disc ratio of this nerve? (Range 0.1-0.9) <input style="width: 80px;" type="text"/></p> <p>*) Please check all features you see in this nerve:</p> <p> ☐ Hemorrhage ☐ Notch ☐ Myopic/tilted disc ☐ Nerve fiber layer defect </p>
--	--	--

Please leave any comments or feedback you have about this HIT or project:

You must ACCEPT the HIT before you can submit the results.

Figure 2.

Screen capture of the authors' Amazon Mechanical Turk (AMT) interface used for grading of fundus images for glaucomatous abnormalities.