Increasing the Bandwidth of Crowdsourced Visual Question Answering to Better Support Blind Users

Walter S. Lasecki*
Dept. of Computer Science
University of Rochester
wlasecki@cs.rochester.edu

Yu Zhong*
Dept. of Computer Science
University of Rochester
zyu@cs.rochester.edu

Jeffrey P. Bigham
HCI and LT Institutes
Carnegie Mellon University
jbigham@cmu.edu

ABSTRACT

Many of the visual questions that blind people ask cannot be easily answered with a single image or a short response, especially when questions are of an exploratory nature, *e.g.* what is in this area, or what tools are available on this work bench? We introduce RegionSpeak to allow blind users to capture large areas of visual information, identify all of the objects within them, and explore their spatial layout with fewer interactions. RegionSpeak helps blind users capture all of the relevant visual information using an interface designed to support stitching multiple images together. We use a parallel crowdsourcing workflow that asks workers to define and describe regions of interest, allowing even complex images to be described quickly. The regions and descriptions are displayed on an auditory touchscreen interface, allowing users to know what is in a scene and how it is laid out.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues – Assistive technologies for persons with disabilities

Keywords

Visual question answering, Crowdsourcing

1. INTRODUCTION AND BACKGROUND

Crowdsourcing answers to visual questions can help blind and low vision users better access the world around them [1]. Prior work on systems such as VizWiz, an application that has allow thousands of blind users to take a picture, speak a question, and get an answer from groups of online workers ("the crowd") in around 30 seconds, and Chorus:View [3], a system that answers questions via streaming video and ongoing conversation between the crowd and users, have shown that crowdsourcing can effectively provide a source of answers to visual questions. In this paper, we present RegionSpeak, a system that builds on the VizWiz platforms and allows blind users to more easily explore a scene by getting more information in a single interaction.

*Equal authorship, listed in alphabetical order.

ASSETS '14 Rochester, NY USA
ACM 978-1-4503-2720-6/14/10.
http://dx.doi.org/10.1145/2661334.2661407.

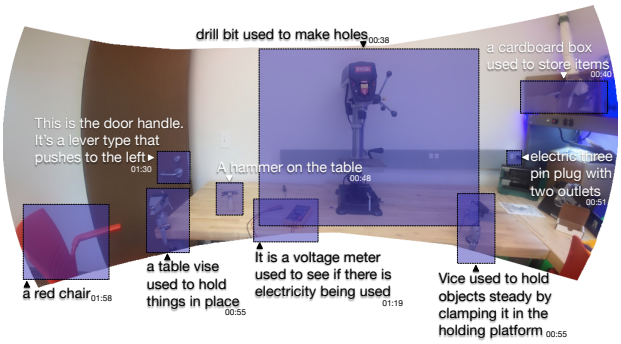

Figure 1: An image of a workbench formed by stitching images together. Important regions are identified and labeled by crowd workers within seconds so that blind users can explore the regions spatially to more easily find what they want.

VizWiz struggles with up to 18% of questions asked because the single-image, single-response model does not efficiently help users frame the information required. Chorus:View overcomes this by engaging users in *continuous* interactions with the crowd via voice and video to help reduce the overhead associated with multi-turn interaction. However, video-based approaches are expensive, more difficult to scale, and can be cumbersome for end users who must actively wait while the crowd determines a response. Our goal is to account for the large set of tasks that fall somewhere between the ideal case for single images in VizWiz, and the continuously-engaged interaction of Chorus:View.

2. IMAGE STITCHING

The first way to reduce the number of interactions needed to answer a question is to allow users to provide the crowd with more information in a single turn. We explore image stitching as a way to do this.

RegionSpeak combines an image stitching algorithm and a key frame extraction algorithm to created a panorama interface for RegionSpeak which has no restriction that needs visual inspection – users can move the camera in any direction, and the key frame extraction algorithm will detect substantial changes in view port and alert users to hold their position to capture a new image. RegionSpeak then takes a photo automatically when the view port is stabilized and gives users audio cue to move on. Users can choose to stop the process, or keep going until they hit the limit of 6 photos. The photos are then sent to our server and stitched before being sent to the crowd.

2.1 Evaluation

We conducted a study with 5 blind people to compare RegionSpeak with single picture approaches (over 28 tasks total). The study was conducted remotely from the blind participants' home using their own iPhones. Participants were paid $10 each, and consented online.

Stitching completed all three tasks within the 10 minute limit, with an average time of 141.1 seconds, while VizWiz failed 1 of the user name reading tasks, and had an average time of 250.5 seconds. This difference is significant, $t(28) = 2.29$ ($p < .01$). The average number of Q&A iterations it takes for stitching (1.79) to yield right answers is also significantly lower than VizWiz(2.79), $t(28) = 2.90$, ($p = .03$). The results confirmed with the stitching interface blind users were likely to capture more visual information in each dialog turn and save time and iterations in subsequent interactions.

In exit interviews, participants said that stitching was easy to understand, learn and use, and they preferred using a stitching interface when taking photos for all task types in our study. They wanted to continue using the stitching interface after the experiments ("*look forward to seeing it released to the general public*"). The feature participants liked most in the stitching interface was the audio guidance which allows "*easy identifying of many things*" and really helps when looking for a specific but small piece of information. It was also mentioned that the fact our stitching interface "*cleverly put different images together figures out the orientation of them*" give them more freedom of interaction.

3. REGIONSPEAK

The other way in which we can reduce the number of interactions required to answer a users question is to allow the crowd to provide more rich answers to users. We introduce RegionSpeak for this purpose. Users begin by either taking a single image or a stitched image, at which point they can set the phone down and wait for responses from the crowd. When responses arrive, RegionSpeak opens up the real time camera view port and starts aligning regions marked by the crowd in the view port using OpenTLD [2]. When a region returned from the crowd is recognized, it is added as an overlay on the camera view port and tracked in real time as the camera is being re-framed. Users can then use their finger to explore the scene, using an interface similar to Voiceover.

RegionSpeak's worker interface asks them to select an important region of the image that they will provide a label. Workers are left to select this region on their own, just as they would be left to choose what level of description to give in a typical VizWiz task. Our selection process is similar to LabelMe [4], which asked crowd workers to carefully select objects and areas in an image by outlining them, but using simple rectangles to make it easier for workers to complete the task quickly, and for end users to find the rough boundaries when scanning the screen with their finger.

3.1 Evaluation

We evaluated RegionSpeak on five images that are similar to questions frequently asked by VizWiz users: a set of five packages of food, a simple diagram on a whiteboard, a set of buttons on a microwave, a menu from a restaurant, and an outdoor scene in a commercial area.

For each of the images, we collected region tags and descriptions from five workers. We coded five features: validity, minimalism, number of objects identified, number of details given, and number of spacial cues provided. For number of objects identified and number of spacial cues, bounding boxes were counted as object identifiers (assuming they contained valid label for a portion of the image). Additional objects and details could be identified within the tag as well. The inter-rater reliability, measured using Cohen's kappa, was between 0.69 and 0.95. We also added "bound tightness" to determine how well workers selected an appropriate region for a given label. We coded all 25 marked segments with two coders. There was strong inter-rater agreement between both raters (Cohen's kappa .74).

Overall, these labels resulted in no minimal answers; an average of 5.2 distinct items marked (median 5, $\sigma = 1.64$); an average of 5.2 descriptive details (median 6, $\sigma = 2.95$); and an average of 6.2 spacial cues (median 5, $\sigma = 2.39$). Additionally, 75% of segments marked by workers were rated as being a "tight bound" on the object they were framing, 20% were considered a "loose bound", and just 5% (1 marking) was rated as an incorrect bound.

However, because the validity of tags was marked per-image, as would be the case with a single description from a worker in our baseline labeling example, just 20% of our images were rated as containing a completely valid label set, with the remaining 80% being rated partially correct. None of the label sets were entirely wrong. This highlights an important aspect of aggregating answers from the crowd: by using aggregated answers, it is more likely that *some* error is introduced, but the chance of an answer containing *entirely* errors falls similarly. In our case, "partially correct" ratings were almost always small errors in one or two labels. On average, responses took 1:05 minutes to arive.

4. CONCLUSIONS AND FUTURE WORK

Our results show that RegionSpeak's image stitching provides a faster and easier means for blind users to capture visual information, and that spatial region labeling encourages crowd workers to provide more descriptive results than traditional labeling. Our next steps are to integrate user feedback into RegionSpeak, and deploy it to the existing VizWiz platform so that users have access to these features.

RegionSpeak fills an important role between existing lightweight visual question answering tools such as VizWiz, which use a single phone image and elicit single responses from workers, and conversational approaches such as Chorus:View, which engage users in longer conversational interactions for questions that require maintaining context across multiple questions. RegionSpeak allows users to send and receive more information with the crowd in each dialogue turn, significantly reducing the number of interactions and the total time spent finding answers.

5. REFERENCES

[1] J. P. Bigham and et al. Vizwiz: Nearly real-time answers to visual questions. In *UIST 2010*.

[2] M. K. . M. J. Kalal, Z. Tracking-learning-detection. In *Pattern Analysis and Machine Intelligence*, 2012.

[3] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham. Answering visual questions with conversational crowd assistants. In *ASSETS 2013*.

[4] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*.