Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search

Darren Guinness
University of Colorado Boulder
Microsoft Research
Boulder CO, USA
darren.guinness@colorado.edu

Edward Cutrell
Microsoft Research
Redmond WA, USA
cutrell@microsoft.com

Meredith Ringel Morris
Microsoft Research
Redmond WA, USA
merrie@microsoft.com

ABSTRACT
Accessing images online is often difficult for users with vision impairments. This population relies on text descriptions of images that vary based on website authors' accessibility practices. Where one author might provide a descriptive caption for an image, another might provide no caption for the same image, leading to inconsistent experiences. In this work, we present the Caption Crawler system, which uses reverse image search to find existing captions on the web and make them accessible to a user's screen reader. We report our system's performance on a set of 481 websites from alexa.com's list of most popular sites to estimate caption coverage and latency, and also report blind and sighted users' ratings of our system's output quality. Finally, we conducted a user study with fourteen screen reader users to examine how the system might be used for personal browsing.

Author Keywords
Accessibility; vision impairment; image captioning; screen readers; alternative text; alt text.

ACM Classification Keywords
K.4.2. [Computers and society]: Social issues – *assistive technologies for persons with disabilities.*

INTRODUCTION
Understanding images on the web can be a difficult task for people with vision impairments. One accessibility practice used to make images more accessible is to describe images using *alternative text*, often abbreviated as *alt text*, and specified within the "alt" attribute of the "img" element in HTML. According to the WCAG 2.0 standard, alt text specifies an HTML property that provides a short text alternative to the image that should contain any words in the image and must convey the overall meaning of the image [36]. However, this practice has not been universally adopted. Prior investigations examining alt text have demonstrated coverage of about 50% of images in heavily trafficked websites [6]. However, alt text coverage is also a feature used to rank pages in search engines [24] and can often be abused, meaning that the alt text might simply state the image's filename or say "img" or "a picture in this page" [11]. This often means that an image on one domain might have a high-quality description, while another site might have a poor-quality description or none at all for the same image, leading to an inconsistent browsing experience for screen reader users.

Researchers have been working to increase online image accessibility through a large range of approaches. The approaches focused on captioning images can be divided into three categories: crowdsourcing, machine-generated, and hybrid systems. Crowdsourcing captioning methods rely on human annotators in order to help describe a photo [4, 8, 9, 27, 34]. Machine-generated captioning approaches rely on machine learning, usually using trained computer vision models to recognize objects and relationships in an image and to produce a caption [12, 13, 17, 26, 33, 37]. Hybrid approaches often rely on a combination of automatically generated descriptions or tags with human editing [6, 28, 29] to reduce time and financial costs associated with recruiting human crowd workers.

Our work contributes to the effort to improve the availability of image captions using a fully-automated approach, but unlike prior automated systems we do not rely on computer vision. Instead, our insight is that many images appear in several places across the web; our Caption Crawler system finds cases where some copies of an image have alt attributes, and propagate this alt text to copies of the image that lack the description, thereby producing human-quality captioning. Our research explores the feasibility of increasing access to shared images online using existing search infrastructure without incurring additional costs in human labeling time or other resources.

This paper presents three main contributions:

1. To provide an understanding of caption coverage in 2017, we present an updated investigation into the alt text coverage on 481 sites sampled from alexa.com's list of the most popular websites [1] in ten categories.

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ACM 978-1-4503-5620-6/18/04 $15.00
https://doi.org/10.1145/3173574.3174092

2. We developed Caption Crawler, a working prototype based on a browser plugin that allows users to browse their favorite websites while image captions from other pages with the same image are dynamically loaded into the browser in the background. In addition, if more than one description is found, users can access the caption queue to gain additional details from other captions retrieved for an image.
3. To test the performance, utility, and usability of our system, we provide results from both automated system tests and user tests with both sighted and visually impaired (VI) users. We found that automated vision-to-language techniques such as CaptionBot [10] performed worse than alt text found on the web, and that our Caption Crawler system was able to significantly reduce the number of uncaptioned images in a page. We also found that users enjoyed having and using the caption queue to gain a broader understanding of image content with additional captions.

RELATED WORK

Image Accessibility on the Web

Ramnath *et al.*'s recent analysis of social image sharing platforms shows a near exponential growth of photo uploading and hosting on the web. The authors highlight that relying on human annotators to caption the web will likely not be able to keep up with the rise in photos online, and propose an automated image captioning system as a scalable solution [26]. Morris *et al.* surveyed blind Twitter users to investigate how their use, needs, and goals with the platform have evolved over the years [22]. This work and studies of blind users on Facebook [35, 37] demonstrate the interest in social media of blind users, but also highlight access issues in the platforms and a trend toward these media becoming increasingly image-centric.

Bigham *et al.* examined both sighted and blind users' browsing activities over the course of a week for a real-world view at the behaviors and accessibility issues related to general browsing. The study demonstrated that about half of all images contained alt text on the original source, and that blind individuals were more likely to click on an image that contained descriptive alt text [3]. In general, studies of consumer and government websites repeatedly find high levels of non-compliance in providing alt text descriptions for online images, with high proportions of descriptions that are missing completely or of very poor quality (*e.g.*, filenames, the word "image") [14, 18, 19, 23, 25].

These prior findings on the increasing pervasiveness of online imagery and the continuing lack of compliance with providing alt text descriptions motivated our development of the Caption Crawler system.

Approaches to Generating Alt Text

Researchers have proposed several approaches to generate alt text for online images. These approaches can be broken into three categories: machine-generated, human-generated, and hybrid approaches.

Machine-Generated Captions

Recent advances in computer vision through deep learning, specifically vision-to-language systems such as [13, 33], have enabled the creation of automatically-generated captions. For instance, the technology described in [13] led to Microsoft's CaptionBot API [10], which generates natural-language descriptions of images. However, studies of the output of vision-to-language systems indicate that they are still too error prone to be reliable for people who are VI [20, 29]. Facebook's Automated Alternative Text system uses a more conservative approach to harnessing the output of computer vision by generating a series of tags that apply to an image with high probability (*e.g.*, "This image may contain the sun, people, the ocean.") [37].

Telleen-Lawton et al. [32] wrote about the use of content-based image search for image retrieval. They presented a series of models along with descriptive user scenarios and made suggestions about the types of performance metrics that were important in each context. Keysers *et al.* developed a system to automatically add alternative text to images based on previously stored images found to be similar. The system used content similarity measures to find a set of visually similar images that contained labels in their database. The system then used these labels to predict possible descriptions for the unlabeled image and could give simple labels such as the color information, as well as the presence of people in the photo and their location [17]. Similarly, Zhang et al [38] used visually similar image results to build a set of words or tags associated with the image. They then used this set of tags in a search query, expanding the original set of tags with commonly used words found in the resulting set of documents after the text search query had been run. The expanded set of tags was then associated with the image and could be used to supply additional context.

Elzer *et al.* created a system that automatically gave descriptions for bar graphs for VI users in their web browsing. This system used computer vision to determine high-level descriptions including the plot title and trends of the data [12].

Human-Generated Captions

Von Ahn *et al.* first introduced the idea of human-powered captioning, using the ESP game to motivate online workers to create tags for images [34]. Takagi et al. [30] demonstrated that readers online could be used to assess and improve accessibility barriers online. Their work presented several tools which allowed users to report accessibility issues and suggest fixes to the website author and other users. Bigham *et al.* showed that crowdsourced image labels could be created in near-real-time (on the order of a few minutes) with their VizWiz system [4], an application that enabled an individual with a visual impairment to take photos of artifacts in their environment and then ask sighted crowd

workers questions about the image. The authors presented a follow-up system VizWiz 2.0 [5] that gave users near real-time feedback to correct issues related to the image and ultimately obtain better answers.

Friendsourcing, a type of crowdsourcing in which social network contacts rather than paid workers perform tasks, has also been investigated as a method of image captioning. Brady *et al.* investigated whether friends of people with VI could be leveraged to answer questions related to inaccessible social media online as a means of lowering the costs of Q&A [8]. However, the study found that that VI users were more reluctant to ask for responses to Q&A than sighted individuals, likely due to associated social costs [8]. This led Brady *et al.* to subsequently introduce the concept of social microvolunteering [9], in which third parties "donate" their social media contacts to label images for VI users in near-real-time; however, it is unclear whether social microvolunteering approaches can scale.

Rodríguez Vázquez investigated the accessibility of images online for non-English speakers [27]. Rodríguez Vázquez posited that online translators, which were already being used to enable multilingual access, could be leveraged to add image descriptions with the proper tools [27].

Zhang *et al.* presented a proxy-based method for runtime repair of Android applications [39]. This allows for inaccessible content to be made accessible at runtime without modifying the original app source. The system enables an end-user or author to update the inaccessible artifact within a proxy with a description or tag that is delivered to future users. The authors demonstrated the system's ability to correct missing accessibility metadata and missing screen reader interactions, modify navigation order, and generate a fully updateable proxy interface.

Hybrid Generation

Neil Rowe proposed Marie-4, which used a web crawler and caption filter to find captions for users when searching for images on the web [28]. The user was actively involved in the retrieval of captions with this system and could search the web with a short keyword search phrase; the system would retrieve images similar to the keywords and captions for the images based on a trained model using HTML elements, and accessibility meta data in the page [28].

Salisbury *et al.* [29] created TweetTalk, a system in which a vision-to-language system provides an initial caption for an image, but then supports a structured dialogue between a user with vision impairments and a crowed worker so that the person with vision impairments could ask clarifying questions about the automated caption. These conversations can then be fed back into machine learning models to improve automated captions for future users.

WebInSight [6] is a system that allows people who are visually impaired to browse websites using a proxy that automatically annotates a page. When a page is requested, the system automatically pulls previous alternative text that has been supplied to its database. This alternative text was generated using three image labeling modules: (1) a Web Context Labeling module, which would retrieve the title and header elements from a page linked to by an image to act as an alternative text description; (2) An Optical Character Recognition (OCR) Image Labeling module which would extract embedded text in images and then perform a spell check and search query on Google using this text to determine if the text was considered "valid"; and (3) a Human Labeling module which would act as a last resort for alternative text descriptions. The authors also mention that the human labeling module had monetary costs, and allowed the user to decide when to use these resources. This system was an inspiration for our own, as it was one of the first automated systems for providing image descriptions. However, this system retrieved other HTML metadata such as using anchors and titles as alternative text rather than existing image captions provided by authors on the web.

Our Approach

Prior human-in-the-loop and hybrid approaches to supplying captions incur costs in money, time, privacy, accuracy, and/or social capital, while fully-automated approaches using computer vision do not yet produce human-quality captions. Our approach attempts to merge the benefits of a fully automated system with the quality of human-authored content by mining existing human-authored captions and applying them to replicated copies of the same image. While this approach cannot caption all images (*e.g.*, unique images that only appear in a single place online), it can increase the accessibility of many online images.

CAPTION CRAWLER: SYSTEM OVERVIEW

Our Caption Crawler system is made up of a client browser extension and a Node.js cloud server. The browser extension searches the DOM of the active webpage for image tags and background images, which are then sent to the server for caption retrieval. When the system finds a caption for an image, the caption is streamed back to the browser extension, which then dynamically appends the caption to the image element. When a user's screen reader focuses on the image, the caption is then spoken aloud to the user. The system prepends the phrase "Auto Alt" before the caption is read so that the user is made aware that the alt text was retrieved automatically by our system, rather than produced by the author of the current web page.

Browser Extension

We developed a Google Chrome Browser Extension to allow dynamic captioning in the user's browser. This extension is broken into two parts: the background script and the content script. The background script is a persistent script that is started upon opening the user's browser. The content script is created every time a new webpage is loaded. The content script uses JavaScript to identify all images in the page, and any captions on the page or accessibility-related metadata from all images and background images in the page. These are then sent to the persistent background script, which transmits this data to the server for alt text retrieval. Alt text

and other captions "scraped" from every browsed page on the client side are also sent back to the server for caching.

Our system is configured to retrieve several types of image captioning including alt text, aria-labels, figcaptions, and elements that include "caption" in their class property. Alt text is an HTML image property which provides a short text alternative to the image [36]. The aria-label attribute is a text description of an image in cases where a caption isn't visibly present in the DOM [21]. Aria-labels are typically used by developers to label buttons when the label of the button is a styled image; however, aria-labels can sometimes be found on images serving as alt text, as they can be applied to any HTML element [21]. Figcaption tags were first defined in HTML 5, and represents a caption or legend associated with a figure [21].

Figure 1: This image is from an online article about volcanos in the arctic [Image credit: NASA / Jim Yungel]. Normally, the screen reader would output this low-quality alt text supplied by the page author, which is the filename, "/h3fs0xudf30016ftkah.jpg". Our Caption Crawler instead produces the alt text, "Auto Alt: Antarctic volcano Mount Erebus seen over the NASA P-3's right wing during the approach to McMurdo Station's sea ice airfield."

Retrieved captions are streamed back to the browser extension, which dynamically appends them to the DOM by replacing the alt text property or adding an aria-image label to a background image. If an image has a poor-quality caption provided by the author, the user can press a key to force a search and overwrite that caption with a new caption from our system. When the screen reader reaches the element, the new caption text will be spoken aloud to the user like any other alt text or aria-label (see Figure 1).

Server

The server comprises a Node.js cloud server on Microsoft's Azure cloud services. The server is written in Node.js and listens for image requests, which are sent to the server via web socket messages. Once a request is received, the server first examines its local databases to determine if the image has cached captions from an earlier request. The database then sends back any cached results, which are streamed into the page while it is still loading. When no cached captions can be found for an image, the system makes a request for a list containing the other pages that display the same image using the Bing Image Insights API [16], which is a part of Azure's Cognitive Services (Google used to offer a reverse image search API, but it is no longer publicly available). We use this API to look up the places online where a specific image is hosted. The API will return a list of URLS that display the image, which are then used to retrieve captions for the image.

When the system receives a list of pages associated with an image, the results are placed in a queue. The system then launches a web crawler for each page in the queue. Each crawler uses 50 simultaneous connections, and uses server-sided JQuery to inspect the DOM for alt text, figcaptions, aria-labels, and other accessibility-related metadata for the matched image on each page. Image matching was performed by using a URI matcher. Our system built a dictionary object for faster lookups for each image request. This dictionary contained all source URLs retrieved from the Image Insights API for the matched image. The matcher would take the path of the image being crawled and compare it with any requested images. If the image was in the request queue, the caption would be automatically placed in the caption queue and made ready for streaming.

When captions are found for an image, they are streamed into the user's browser extension via a web socket connection to the server. The browser extension then dynamically adds the caption to the page in the form of alt text for image elements and aria-labels for background images.

Our system also extracts the alt text and image captions in the DOM while the user is browsing a page using the browser extension. This allows the system to keep improving as more pages are browsed by users.

If we find multiple potential captions for the target image, we stream the longest caption by default, since that performed best in our caption-quality rating evaluation (described later in the "Caption Quality and Preference Ratings" section). However, we keep a queue of all captions found; if the user is not satisfied with a caption, they can press a shortcut key to access additional captions from the queue.

Caption Crawler automatically supplies captions when alt text is missing entirely; however, sometimes images contain alt text, but it is of poor quality. If a user would like our system to replace a poor-quality alt text, they can press a keyboard shortcut to request a replacement of the alt text with a caption from our system. When the screen reader observes the alt change, it automatically speaks the new caption.

If our Caption Crawler is unable to find a pre-existing caption for an image on the web, it requests a computer-generated caption from Microsoft's CaptionBot API [10], which uses computer vision to describe an image. When the text from CaptionBot is read aloud, the screen reader first speaks the words "CaptionBot:" so that the user is aware that this is not a human-authored caption.

SYSTEM PERFORMANCE

To gain an understanding of the performance of our Caption Crawler system, and to collect metadata around images on commonly visited websites, we performed a crawl of 500 pages from popular websites using our system. This crawl ran over the top 25 and bottom 25 websites over 10 categories taken from Alexa's list of the most popular websites [1] (i.e., 50 websites in each of 10 categories). The categories used were: Arts, Business, Health, Kids & Teens, News, Reference, Science, Shopping, Society, and Sports. The caption-related statistics for each category can be seen in Table 1. Each page was crawled for one minute to gauge the number of captions we could retrieve within that time limit, since we anticipate that users would not tolerate a high-latency system.

We wanted to examine the worst-case performance of our system for the data crawl, so we disabled the database cache. Instead we requested the list of URLs displaying the same image through the Image Insights API and launched web crawlers for each successful query to get a better understanding of how the system would handle new pages. We also did not use CaptionBot to supply missing captions during this crawl, as our focus was on understanding the extent to which pre-existing captions can provide coverage of missing content.

Alexa Crawl Results

We collected metrics during our data crawl, which can be seen in Tables 1 and 2. Table 2 displays the breakdown of images and background images found without alt text, or captioning, and the number of captions our system was able to retrieve, as well as the caption coverage with and without the system.

Our system was able to crawl 481 of the 500 pages in the set; one website was removed for adult content, while the other eighteen had excessive JavaScript reloading in the browser, which would cause the statistical entry for the page to fail to upload. For these 481 pages, a total of 8,435 unique images lacked alt text; our system was able to stream into the page pre-existing alt texts from other websites for 1,013 of these images (12%). Our system retrieved a total of 11,492 captions during the crawl; we were only able to stream in 1,013 captions because some captions were collected after the one minute timeout period, and because many captions were often found for a single image, in which case we only streamed the longest (any extra captions were then added to that image's caption queue). For these images that had pre-existing alt text, we typically found multiple possible alternatives, which we kept in our queue.

Captions requested	Captions added	Retrieved captions (server)	Scraped captions (client)	Avg latency (sec)
8,435	1,013	11,492	3,728	18.11

Table 1. Overall Metrics gathered from the data crawl (all images and captions are unique). "Captions requested" is the number of images sent out for captioning during the crawl. "Captions added" is the number of captions streamed into the page by our system. "Captions retrieved" is the total number of captions collected by the crawler. "Scraped captions" represents the number of captions extracted by the client's browser and sent to the database for caching. "Average latency" represents the amount of time between requesting a caption and receiving a result.

For images for which our system was able to retrieve at least one caption, we found an average of 1.82 existing captions (SD = 6.31). The average latency for finding all available captions for images in a page was 18.11 seconds (SD = 15.35 seconds). As shown in Table 2, the alt text coverage of our system varied by page category, ranging from a low of

	Images no alt	Alt added	BG images	Aria added	Total Images	Alt coverage old (%)	Alt coverage new (%)	Alt Added (%)
Arts	376	84	564	65	1270	70.39	77.01	22.34
Business	258	31	535	44	1076	76.02	78.90	12.02
Health	269	36	421	42	1151	76.63	79.76	13.38
Kids	311	25	496	42	921	66.23	68.95	8.04
News	559	80	433	46	1835	69.54	73.90	14.31
Reference	89	7	186	12	441	79.82	81.41	7.87
Science	179	31	491	45	971	80.05	83.86	17.32
Shopping	282	71	652	48	1695	83.36	87.55	25.17
Society	284	30	334	32	822	65.45	69.10	10.56
Sports	398	33	539	76	1312	69.66	72.18	8.30
U.S.	430	81	407	52	1149	62.58	69.63	18.84
User Study	935	127	667	21	2680	65.11	69.85	13.58
Average	364.16	53	477.08	43.75	1276.92	72.07	76.08	14.31

Table 2. Comparison of unique image metrics including caption coverage before and after using the system in the Alexa data crawl categories (first 10 rows) and user study. Alt Added is the percent of blank alt tags that have been updated by the system.

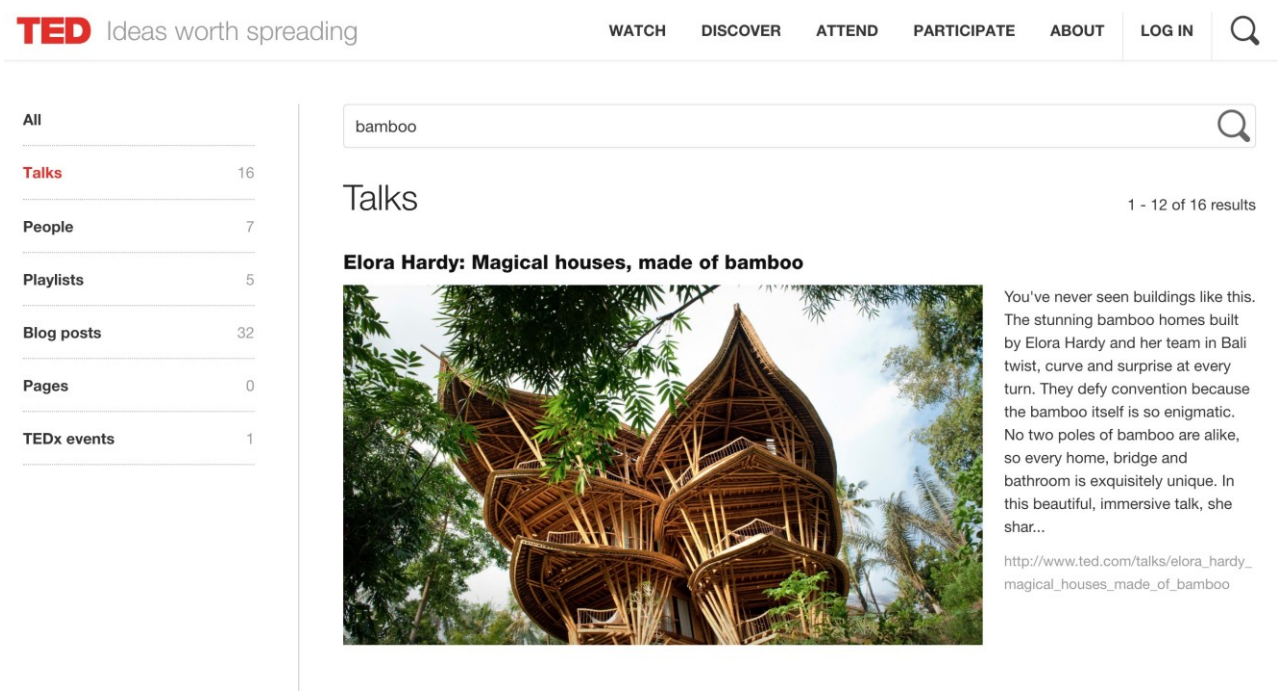

Figure 2. Image of the TED website featuring an image without alt text or a caption. The featured image of a grand bamboo home in Bali, pictured from below was used in our caption quality questionnaire. Retrieved captions for this image were: *First:* Fantasy Bamboo Tropical House; *Longest:* Sharma Springs is a six-storey bamboo house constructed by Ibuku in Bali; *Random:* Elora Hardy ibuku bamboo; *Bag of Words:* Image may contain: bali, bamboo, houses, ibuku, sharma; *Bing Caption:* In this bamboo tree house in Bali, they've implemented a bamboo bridge to enter the dwelling; *Caption Bot:* I really can't describe the picture.

finding around 8% of missing alt texts for pages in the "Kids" and "Sports" categories, to a high of finding around 25% of missing alt texts in the "Shopping" category.

To determine if our system improved image caption coverage in the browser, we ran a paired t-test over the number of uncaptioned unique images present with and without the system over each of the ten alexa.com categories. The t-test ($t(10) = 5.65$, $p < .001$) identified that the system significantly improved caption coverage at the $p < .001$ level. Similarly, another paired t-test ($t(10) = 9.27$, $p < .001$) demonstrated the ability for the system to significantly improve caption coverage for background images in browsing.

In addition to demonstrating the performance of our system, these findings also demonstrate that alt text compliance continues to be incomplete, even on popular sites.

CAPTION QUALTY AND PREFERENCE RATINGS

If our Caption Crawler identifies several possible captions for a single image, we must choose which to supply as the default replacement. We considered several possible methods for making this selection (Figure 2), including:

- *first* (use the first caption found, for speed)
- *longest* (select the longest caption found, since length has been found to correspond to quality in other domains, such as Q&A systems [15, 31])
- *random* (randomly choosing from the options in our caption queue)
- *bag of words* (creating a caption jointly from all items in the queue by extracting unigrams from all captions in our queue, removing common stop words, and then reporting the unigrams most common among the set)
- *Bing caption* (using a caption returned by the Bing Image Insights API, which uses a natural language processing model to summarize the text surrounding the image in the page to be used as a caption)
- *CaptionBot* (using a vision-to-language API to create an automated caption)

To help us better understand how well the different captions generated by the Caption Crawler described images from the

	%First (sighted)	%First (VI)	Mean Rank (sighted)
First	18%	15%	3.41
Longest	24%	27%	3.16
Random	19%	11%	3.38
Bag of Words	9%	11%	3.39
Bing Caption	19%	26%	3.27
CaptionBot	11%	11%	4.40

Table 3. Percent of each retrieval type as ranked the best (1) between sighted and visually impaired (VI) participants, and mean rank (sighted) over each category (1 is best, 6 is worst).

web, we created two online questionnaires. In the first questionnaire, we presented sighted people with fifteen images taken from our Alexa crawl. For each image, we presented six different captions generated as described above (first, longest, random, bag of words, Bing caption, and CaptionBot). Respondents were asked first to identify which caption best described the image, and then to rank order the remaining captions from best to worst

While the first questionnaire was useful in understanding how sighted people rated the quality of image descriptions, we were also interested in how the descriptions provided by the system would be judged by respondents with visual impairments. To this end, we generated a second questionnaire that we sent to people with visual impairments who regularly use a screen reader. For this survey, we asked respondents to listen to the same captions presented in the first survey: six captions for fifteen different images. For this group, we only asked respondents to identify the best captions for each image and did not request a ranking of the other, non-preferred captions (this was to keep the survey length reasonable, as it takes longer to complete when using a screen reader).

For the first questionnaire, we recruited 40 sighted respondents from our organization (25 female) with a mean age of 22. All sighted participants were currently students, ranging from undergraduate to Ph.D. candidates. For the second questionnaire, we used email lists from organizations related to visual impairment to recruit 39 respondents with vision impairments who use a screen reader (20 totally blind, and 19 with very low vision).

For both sighted and visually-impaired respondents, the longest caption was most frequently identified as best, though this was not uniformly true for all images or judges (see Table 3). For the first survey, where we had full ranked data, we performed a Friedman test of the mean ranks. The test was significant, $\chi^2(5, N=600) = 173$, $p<<0.001$. Kendall's W=0.058, indicating weak differences between the six types of captions. Follow-up pairwise comparisons were mostly uninteresting, with the only significant differences between CaptionBot and each of the other types, reflecting prior findings that vision-to-language systems do not yet equal human-quality captions [20, 29]. Based on these findings, we set the longest caption to be the default used by our system.

USER STUDY

While our crawl of nearly 500 popular sites gives information about our system's performance on popular web pages, we wanted to understand how our system would do on the pages important to VI users, which may overlap with popular sites, but may also represent pages from the "long tail" of popularity distributions. We also wanted to understand end-users' reaction to the concept of substituting alt text from one web page into another, and of the ability to interact with our caption queue to sample multiple alt text alternatives. To address these goals, we conducted a user study in which screen reader users came to our lab and used the Caption Crawler system to browse sites they would typically visit.

Participants

We recruited fourteen legally blind adults from the Seattle metropolitan area using email lists for local organizations related to visual impairment. Participants spent thirty minutes at our lab completing the study and were paid for their time ($25 for thirty minutes plus transportation costs). Participants' ages ranged from 25 to 65 years old (mean = 43.7 years), and half were female. Eight participants described themselves as either completely blind or having only light perception, while six had very small amounts of residual vision; all participants used a guide dog or white cane, and all relied on screen reader technology when interacting with computers.

Method

To better understand how the Caption Crawler system would perform *in situ*, we engaged our participants in a web browsing session. Sessions were conducted on our lab's computer, which ran Google Chrome on the Windows 10 operating system and read web pages aloud using the Windows Narrator screen reader. After a brief tutorial in which participants heard our system provide alt texts for an example site, participants suggested pages that they were interested in viewing. Note that we told participants not to suggest personal social media accounts (*e.g.*, Facebook) for viewing with our system, due to privacy concerns around accessing their friends' images. With the assistance of the researcher in the room, they visited each suggested page with a focus on the images. Participants were presented with the image descriptions in the page and could request additional captions from the queue for any image on the page.

Participants heard the system's output for between two to six of their chosen sites in the thirty-minute study session (we continued to visit sites until running out of time, hence the variation in how many sites each participant was a ble to experience). A total of 72 pages (67 unique) were browsed by our participants for at least one minute during the user study. Thirteen of these unique pages (19%) overlapped with the Alexa crawl. Example pages visited included shopping pages for specific products of interest to participants (*e.g.*, searching for cookware on Amazon, searching for clothing on Lands' End), blogs relating to participants' hobbies (*e.g.*, a site about safari camps, a blog about a participant's urban neighborhood), and specialized or local news outlets (*e.g.*, local weather information, local sports team pages, a religious news site).

After experiencing each site, we asked participants the following questions, using a Likert scale from 1 to 5, where 1 indicated strongly disagree, and 5 indicated strongly agree:

1. The image descriptions on this page were useful for understanding the content of the image.

2. The image descriptions were clear and made sense to me.
3. Other image descriptions in the queue were useful.
4. If more image description alternatives were available, I would listen to them.

RESULTS

Our system collected several metrics while the user study was conducted, allowing us to perform the same coverage and latency checks as in our Alexa crawl. These metrics can be seen in the last row of Table 2. Our system successfully retrieved pre-existing (not CaptionBot) alt text for 13.58% of images that had lacked it on the set of sites suggested by user study participants; the mean latency was 20 seconds, with a standard deviation of 7 seconds.

To determine if our system had a significant effect on the number of uncaptioned images in the sites viewed by our participants, we again performed paired t-tests. The data used was the number of uncaptioned images present with and without the system over each of the 72 websites visited during the user study. A paired t-test indicated that Caption Crawler significantly improved caption coverage for standard HTML (non-background, non-decorative) images ($t(72) = -5.60$, $p<.001$). Similarly, another paired t-test for background images showed a significant improvement in caption coverage ($t(10) = -3.88$, $p<.001$).

Participants' reactions to the system varied. While all participants were excited by the idea of the system, some participants wanted more details than the Caption Crawler could provide for some images. For example, P9 wanted very detailed image descriptions for all their browsed images. However, P9 did mention that "[the system] gives you something to go on," even when the retrieved alt texts were not as detailed as they had hoped. P6 felt similarly, and stated "Well I guess it does give a little information" but ultimately wanted more detail in the descriptions. P3 spoke about the inconsistency of the retrieved text, saying "the level of detail [in the auto alts retrieved] varied a lot." This is a known issue in our system, which is not designed to supply an alt text for every single image on the web, but rather to efficiently, cheaply, and automatically make strides toward improving accessibility by captioning the subset of images that are replicated online.

Other participants were more positive about the system; for instance, P11 mentioned appreciating the low latency: "I like the fact that it was able to describe photos taken as we speak, it's happening now and instant." P2 wanted to continue using the system after the study and said, "when is it going to be in a product, I'm ready for it." P8 mentioned that they were able to get more out of browsing with the system on. When P7 was asked whether the system was useful compared to her normal browsing experience she responded "Absolutely. Very useful… most, if not all, images I see online don't have good descriptions… this is definitely useful."

	Caption Useful	Caption Clear	Queue Useful	Would Listen
Median	3	2.88	3.33	4.75
Mean	2.82 (0.76)	2.68 (0.58)	3.13 (1.13)	4.15 (1.23)

Table 4. Median and mean Likert ratings, standard deviations in parentheses; 1= strongly disagree, 5 = strongly agree.

Some participants mentioned that they appreciated the ability to replace poor-quality captions (in addition to supplying completely missing captions), and liked being able to use the caption queue to build a better idea of the content in an image by sampling multiple descriptions. P1 mentioned, "it's a good example where the auto alt… actually provided more information than the initial text [provided by the page author]" when speaking about replacing a less descriptive caption with a new one from the system. P1 also liked the ability to use the caption queue, stating, "at least a couple [alternatives are nice to hear] to corroborate what's there." P7 similarly stated, "I think having the ability to go through [the queue] gives some sort of verification… builds some confidence."

Although our surveys already found that both sighted and visually-impaired participants preferred the other caption types over the CaptionBot description, several participants reiterated this outcome during the study. For example, P13 mentioned that "The auto alt is way better than the caption bot." Several other participants wanted to be able to request a CaptionBot description using a different key than that used for retrieving the queue of additional pre-existing alt texts, in order to keep separate mental models of the human-generated versus machine-generated descriptions.

Table 4 shows the average ratings for the Likert questions asked after users experienced a page. We see that users found room for improvement in the system's performance (low ratings typically reflected instances where no pre-existing alt could be found and only a CaptionBot caption was provided); however, users liked the idea of the caption queue, and were eager to spend time listening to alternative captions if more could be retrieved.

DISCUSSION

Our studies show positive trends in alt text coverage since Bigham *et al.*'s study in 2006 that found about half of images lacked alt text [6]; our crawl of 481 popular sites found that on average 72% of images had alt text, which means that web accessibility practices have become more widely adopted, at least on popular websites, though a substantial amount of images continue to lack captions. However, our user study demonstrated that in practice many of the original captions on the page needed to be replaced with additional captions from the queue to allow a user to better understand the content in an image.

While our system was able to provide some high-quality captioning, the consistency of the captioning varied and participants in our user study often used other items in the

caption queue to build a better mental model of what was in the image. At the time of the user study, the Caption Crawler did not produce caption credits, but occasionally the returned caption text contained a credit. However, during the study participants specifically noted that they liked hearing the photo credit and caption credit text on captions from prominent outlets such as Getty Images and used these as a sign of an accurate caption, suggesting that adding sourcing information to all Caption Crawler captions may be desirable to help users build better mental models, and allow end user customization over caption priority.

Our system also occasionally retrieved captions in different languages. Non-English Latin-based texts were announced to users but were not intelligible due to the language pronunciation setting on the screen reader. Captions including other non-Latin alphabets were not announced, as the screen reader was not configured to speak these alphabets. While this was an issue during the user study, we believe that this may ultimately be a feature for non-English screen readers if configured properly (and non-English captions could be filtered for English speakers using machine learning approaches and/or heuristics based on Unicode character sets).

Our system did not attempt to judge caption accuracy or descriptiveness beyond using the heuristic retrieval methods (*i.e.,* prioritizing the longest caption). However, a caption selection model similar to the one used in [2] could be used to reduce the inconsistency of retrieved captions using features such as domain, presence of caption credit, length, and others.

After the user study tasks ended, we asked participants about potential improvements for future versions of the software. One participant (P3) suggested that we attempt to describe logos, which went against our initial design decisions that focused only on captioning content images and ignoring decorative images like logos and menu items. To provide more insight about when the system has helpful information, P1 suggested that we use audio cues or Earcons [7] to allow a user to know if helpful captions have been discovered for an image.

Limitations

Our approach only works on images that have been used in multiple places. While this is true of many images on the web, our system underperforms on unique images that aren't hosted in many locations, such as those in personal photos. An extension that future research could consider is using computer vision or other AI techniques to identify images that are not an exact match for the target image, but which are nonetheless highly similar, in order to expand the set of images for which captions can be retrieved. Our system also has difficulties with encrypted URLs used on sites such as Google, as these encrypted URLs do not show up in image indexing. In the future, we could use the image binaries and hashing to resolve encrypted image matching. Finally, our streamed captions had higher latencies than we aimed for. After investigating the Azure logs, we realized that our multiple crawler approach was reaching the memory limits available in the node VM. This effectively made our caption search a depth-first search rather than breadth-first, filling in the caption queue for a single image before covering the other images on the page due to memory management. Future work should leverage a breadth-first search in which the system aims to add a single caption to each image on the page before fleshing out extra captions for the queues, to achieve a lower average latency.

Design Implications for Search Engines

Our system has the potential to improve its latency as it receives additional requests due to the caching of caption requests. However, we would like to encourage search engines to extract human-crafted captions and alt text during their indexing crawls to reduce the need for systems like Caption Crawler to index all images on the web.

In our testing and piloting we found that some websites contain generic alt text that doesn't contain meaningful image descriptions such simply containing "alt", "img/image," or a filename as an image description. We believe this to be an artifact of authors following Search Engine Optimization (SEO) policies that give weight to images with captions, but that do not weigh caption accuracy. While we have configured our system to choose longer image descriptions and to avoid common inaccurate descriptions, we believe that ultimately SEO policy could incorporate image caption accuracy. If a search engine were to randomly sample an image and caption on a page for accuracy using crowdsourcing, authors might provide better captioning to their entire domain to improve their ranking.

CONCLUSION

In this work, we presented the Caption Crawler system, a real-time system for image captioning that relies on reverse image search to reuse pre-existing captions for the target image from other websites. We performed an automated crawl using the system over 481 web pages selected from alexa.com's list of the most popular websites in order to obtain an idea of the caption coverage on common sites and characterize the performance of our system; we were able to provide alt text for about 12% of images that lacked it, with only 18 seconds' latency on average. We also presented a study of caption preferences that found both sighted and blind users preferred hearing the longest available caption by default, and a lab study with fourteen blind users to further evaluate our system's performance. Our findings indicate that users are receptive to the idea of quickly and automatically replacing missing and low-quality alt texts with pre-existing captions from elsewhere on the web, that pre-existing human-authored captions are preferred to current automated vision-to-language approaches, and that providing a queue of alternative possibilities was valued by end users as a way for them to learn more about an image and have increased confidence in caption accuracy. These performance and user-study findings suggest that our

approach can play an important role in helping to address the issue of missing and poor-quality alt text online.

AKNOWLEDGEMENTS

We thank our participants, and the MSR Nexus team for their feedback and assistance. We would also like to thank Cole Gleason, Cindy Bennett, Martez Mott, Jazette Johnson, and Alex Fiannaca for their helpful feedback.

REFERENCES

1. Alexa top 500 global sites on the web, 2017. https://www.alexa.com/topsites.
2. Bigham, J. P. (2007, January). Increasing web accessibility by automatically judging alternative text quality. In Proceedings of the 12th international conference on Intelligent user interfaces (pp. 349-352). ACM.
3. Bigham, J. P., Cavender, A. C., Brudvik, J. T., Wobbrock, J. O., & Ladner, R. E. (2007, October). WebinSitu: a comparative analysis of blind and sighted browsing behavior. In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (pp. 51-58). ACM.
4. Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., ... & Yeh, T. (2010, October). VizWiz: nearly real-time answers to visual questions. In Proceedings of the 23nd annual ACM symposium on User interface software and technology (pp. 333-342). ACM.
5. Bigham, J. P., Jayant, C., Miller, A., White, B., & Yeh, T. (2010, June). VizWiz:: LocateIt-enabling blind people to locate objects in their environment. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 65-72). IEEE.
6. Bigham, J. P., Kaminsky, R. S., Ladner, R. E., Danielsson, O. M., & Hempton, G. L. (2006, October). WebInSight:: making web images accessible. In Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility (pp. 181-188). ACM.
7. Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. Human–Computer Interaction, 4(1), 11-44. Blattner, Meera M., Denise A. Sumikawa, and Robert M. Greenberg. "Earcons and icons: Their structure and common design principles." *Human–Computer Interaction* 4, no. 1 (1989): 11-44.
8. Brady, E. L., Zhong, Y., Morris, M. R., & Bigham, J. P. (2013, February). Investigating the appropriateness of social network question asking as a resource for blind users. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 1225-1236). ACM.
9. Brady, E., Morris, M.R., and Bigham, J.P. Gauging Receptiveness to Social Microvolunteering. *Proceedings of CHI 2015*.
10. CaptionBot – For pictures worth the thousand words, 2017. https://www.captionbot.ai.
11. Diaper, D., & Worman, L. (2004). Two Falls out of Three in the Automated Accessibility Assessment of World Wide Web Sites: A-Prompt vs. Bobby. In People and Computers XVII—Designing for Society (pp. 349-363). Springer, London.Dan Diaper, and Lindzy Worman. 2004. Two Falls out of Three in the Automated Accessibility Assessment of World Wide Web Sites: A-Prompt vs. Bobby. In *People and Computers XVII—Designing for Society* (pp. 349-363). Springer, London.
12. Elzer, S., Schwartz, E., Carberry, S., Chester, D., Demir, S., & Wu, P. (2007, March). A Browser Extension for Providing Visually Impaired Users Access to the Content of Bar Charts on the Web. In WEBIST (2) (pp. 59-66).Elzer, Stephanie, Edward Schwartz, Sandra Carberry, Daniel Chester, Seniz Demir, and Peng Wu. "A Browser Extension for Providing Visually Impaired Users Access to the Content of Bar Charts on the Web." In *WEBIST (2)*, pp. 59-66. 2007.
13. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., and Zweig, G. From captions to visual concepts and back. *Proceedings of CVPR 2015*.
14. Goodwin, M., Susar, D., Nietzio, A., Snaprud, M., and Jensen, C.S. 2011. Global web accessibility analysis of national government portals and ministry web sites. *Journal of Information Technology and Politics*, 8(1), 41-67.
15. Harper, F.M., Raban, D., Rafaeli, S., and Konstan, J.A. Predictors of Answer Quality in Online Q&A Sites. Proceedings of CHI 2008, 865-874.
16. Image Insights | Microsoft Developer Network https://msdn.microsoft.com/en-us/library/mt712790(v=bsynd.50).aspx
17. Keysers, D., Renn, M., & Breuel, T. M. (2007, October). Improving accessibility of HTML documents bygenerating image-tags in a proxy. In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (pp. 249-250). ACM.
18. LaBarre, S.C. 2007. ABA Resolution and Report on Website Accessibility. *Mental and Physical Disability Law Reporter.* 31(4), 504-507.
19. Loiacono, E.T., Romano, N.C., and McCoy, S. 2009. The state of corporate website accessibility. *Communications of the ACM,* 52(9), September 2009, 128-132.

20. MacLeod, H., Bennett, C. L., Morris, M. R., & Cutrell, E. (2017, May). Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 5988-5999). ACM.
21. MDN Docs - Figcaption.& Aria-label | Mozilla Developer Network https://developer.mozilla.org/en-US/docs/Web/HTML/Element/figcaption https://developer.mozilla.org/en-US/docs/Web/Accessibility/ARIA/ARIA_Techniques/Using_the_aria-label_attribute
22. Morris, M. R., Zolyomi, A., Yao, C., Bahram, S., Bigham, J. P., & Kane, S. K. (2016, May). With most of it being pictures now, I rarely use it: Understanding Twitter's Evolving Accessibility to Blind Users. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5506-5516). ACM.
23. Olalere, A. and Lazar, J. 2011. Accessibility of U.S. Federal Government Home Pages: Section 508 Compliance and Site Accessibility Statements. *Government Information Quarterly*, 28(3), 303-309.
24. Patil Swati, P., Pawar, B. V., & Patil Ajay, S. (2013). Search Engine Optimization: A Study. Research Journal of Computer and Information Technology Sciences, 1(1), 10-13.Patil Swati, P., Pawar, B.V. and Patil Ajay, S., 2013. Search Engine Optimization: A Study. *Research Journal of Computer and Information Technology Sciences*, *1*(1), pp.10-13.
25. Power, C., Freire, A., Petrie, H., and Swallow, D. Guidelines are only half of the story: Accessibility problems encountered by blind users on the web. *Proceedings of CHI 2012*.
26. Ramnath, K., Baker, S., Vanderwende, L., El-Saban, M., Sinha, S. N., Kannan, A., ... & Bergamo, A. (2014, March). Autocaption: Automatic caption generation for personal photos. In Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on (pp. 1050-1057). IEEE. Krishnan Ramnath, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta N. Sinha, Anitha Kannan, Noran Hassan, and Michel Galley, 2014, March. Autocaption: Automatic caption generation for personal photos. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on* (pp. 1050-1057). IEEE.
27. Rodríguez Vázquez, S. (2016, April). Measuring the impact of automated evaluation tools on alternative text quality: a web translation study. In Proceedings of the 13th Web for All Conference (p. 32). ACM.
28. Rowe, N. C. (2002). Marie-4: A high-recall, self-improving web crawler that finds images using captions. IEEE Intelligent Systems, 17(4), 8-14.
29. Salisbury, E., Kamar, E., and Morris, M.R. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Proceedings of HCOMP 2017.*
30. Takagi, H., Kawanaka, S., Kobayashi, M., Itoh, T., & Asakawa, C. (2008, October). Social accessibility: achieving accessibility through collaborative metadata authoring. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility* (pp. 193-200). ACM.
31. Teevan, J., Morris, M.R., and Panovich, K. Factors Affecting Response Quantity, Quality, and Speed for Questions Asked via Social Network Status Messages. Proceedings of ICWSM 2011.
32. Telleen-Lawton, D., Chang, E. Y., Cheng, K. T., & Chang, C. W. B. (2006, January). On usage models of content-based image search, filtering, and annotation. In *Internet Imaging VII*(Vol. 6061, p. 606102). International Society for Optics and Photonics.
33. Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., and Sienkiewicz, C. Rich Image Captioning in the Wild. *Proceedings of CVPR 2016*.
34. von Ahn, L., Ginosar, S., Kedia, M., Liu, R., and Blum, M. Improving accessibility of the web with a computer game. *Proceedings of CHI 2006.*
35. Voykinska, V., Azenkot, S., Wu, S., and Leshed, G. How Blind People Interact with Visual Content on Social Networking Services. *Proceedings of CSCW 2016*.
36. Web Content Accessibility Guidelines 2.0, W3C World Wide Web Consortium Recommendation 05 September 2017. (http://www.w3.org/TR/200X/REC-WCAG20-20081211/)
37. Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017, February). Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1180-1192). ACM.
38. Zhang, X., Li, Z., & Chao, W. (2013). Improving image tags by exploiting web search results. *Multimedia tools and applications*, *62*(3), 601-631.
39. Zhang, X., Ross, A. S., Caspi, A., Fogarty, J., & Wobbrock, J. O. (2017, May). Interaction Proxies for Runtime Repair and Enhancement of Mobile Application Accessibility. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 6024-6037). ACM.