Twitter A11y: A Browser Extension to Make Twitter Images Accessible

Cole Gleason¹, Amy Pavel¹, Emma McCamey², Christina Low³, Patrick Carrington¹, Kris M. Kitani¹, Jeffrey P. Bigham¹

¹Carnegie Mellon University, Pittsburgh, USA
²Virginia Commonwealth University, Richmond, USA
³Stony Brook University, Stony Brook, USA

cgleason,apavel,pcarrington, mccameyec@vcu.edu
chlow@cs.stonybrook.edu
kkitani,jbigham@cs.cmu.edu

ABSTRACT
Social media platforms are integral to public and private discourse, but are becoming less accessible to people with vision impairments due to an increase in user-posted images. Some platforms (e.g. Twitter) let users add image descriptions (alternative text), but only 0.1% of images include these. To address this accessibility barrier, we created Twitter A11y, a browser extension to add alternative text on Twitter using six methods. For example, screenshots of text are common, so we detect textual images, and create alternative text using optical character recognition. Twitter A11y also leverages services to automatically generate alternative text or reuse them from across the web. We compare the coverage and quality of Twitter A11y’s six alt-text strategies by evaluating the timelines of 50 self-identified blind Twitter users. We find that Twitter A11y increases alt-text coverage from 7.6% to 78.5%, before crowdsourcing descriptions for the remaining images. We estimate that 57.5% of returned descriptions are high-quality. We then report on the experiences of 10 participants with visual impairments using the tool during a week-long deployment. Twitter A11y increases access to social media platforms for people with visual impairments by providing high-quality automatic descriptions for user-posted images.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI'20, April 25–30, 2020, Honolulu, HI, USA
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-6708-0/20/04. . . $15.00
DOI: https://doi.org/10.1145/3313831.3376728
share information about their disabilities or organize around disability activism. For example, the #HandsOffMyADA and #CripTheVote campaigns on Twitter were organized by disability activists around pending legislation in the US [1,7]. Auxier et al. found that even in the #HandsOffMyADA campaign, only 7% of images contained alternative text, leaving most of the images inaccessible to people with vision impairments.

To provide high-quality descriptions for images on social media platforms, we designed an end-to-end system, Twitter A11y, to generate or retrieve alt text for images on Twitter (Figure 1). Prior research has developed several automatic and human-in-the-loop methods to generate image descriptions, and these methods are now robust enough to deploy at scale to address the growing lack of image accessibility on social media. Twitter A11y includes three methods that automatically add alt text for user-posted images: text recognition (optical character recognition), scene description, and the Caption Crawler method [11] (reverse image search). Two additional methods seek to address Twitter-specific images categories: screenshots of tweets and preview images for external links. Finally, if none of the prior methods produce a satisfactory alt text description for the image, Twitter A11y asks a crowd worker to describe the image on Amazon Mechanical Turk using a set of provided guidelines. Twitter A11y’s browser extension dynamically requests alt text in the background for images as a user uses the Twitter website, and adds it to the image as if it had been there originally.

To evaluate the coverage and quality of alt text from the six methods, we performed a static analysis of images from 50 blind Twitter users’ timelines. We randomly sampled tweets they may have read over the course of a day, creating a sample of 1,198 images. Through a combination of automatic methods, Twitter A11y increased the alt text coverage from 7.6% to 78.5%. We then rated a subset of these images to compare the quality of the descriptions returned from each method on a four-point scale. We consider the alt text that achieves either the highest rating (“Great”) or second-highest rating (“Good”) to be high-quality alt text. The highest percentage of quality alt text was from the text recognition (32.7% “Good”, 44.9% “Great”) and scene description (53.1% “Good”, 14.3% “Great”) methods. We also evaluated crowdsourcing as an additional method, finding that 62.5% of the resulting descriptions were rated “Great” (and 18.8% “Good”).

We recruited 10 participants who access Twitter via a screen reader, to evaluate the perception of Twitter A11y and the six methods. Twitter A11y was able to add automatic descriptions to 82.4% of the content they accessed, crowdsourcing the remaining 17.6%. On average, participants’ perceptions were that 12.1% of images were accessible in their timelines before the study, and 72.3% of images were accessible when using Twitter A11y.

In this work, we make social media content accessible by asking people with vision impairments to install a browser extension, but the technological and financial costs of making social media accessible should not be borne solely by people with disabilities in the long term. Rather, the platforms should bear the responsibility to ensure their hosted content is accessible through more accessibility features, user education, and employing the methods used by Twitter A11y. Therefore, the Twitter A11y approach and user evaluation results should be informative for application developers, not used as a justification for user-installed solutions.

This work represents a combination and comparison of known methods that are now robust enough to address the accessibility issues plaguing social media platforms. We show the potential for dramatic improvement in accessibility, the differences between coverage and quality of different methods, and the impact of this tool on the social media experiences of our participants. This work opens future directions for researchers to improve and combine the methods used by Twitter A11y and provides guidance for social media designers on integrating methods to make their platforms accessible at scale.

RELATED WORK

This research is related to prior work on (1) automatically generating alternative text for people with vision impairments, (2) crowdsourcing image descriptions, and (3) accessibility on social media platforms.

Automatically Generating Alternative Text

Alternative text is a property in the HTML specification [2] to include a textual description of visual content in an image. It is primarily used by people with vision impairments who browse the web with screen readers or braille displays, but can also be used by search engines [14] and other machine learning applications [13]. While alt text is currently just textual, researchers have proposed updating the standard to include richer representations of images, such as background audio or sound effects [9,19]. The lack of alternative text is a perennial problem on the web [3, 12], and various methods can be used to describe visual aspects of the image.

Optical character recognition attempts to extract text characters captured in images and correct errors to make coherent words or sentences [3]. Object recognition algorithms can locate and identify entities in the image that the model has been trained to recognize, such people or animals [27]. Instead of a list of objects, scene description methods generate a caption for the image, attempting to describe aspects of the image in a grammatically-correct sentence structure. This approach is available in commercial applications like Microsoft Seeing AI [17]. MacLeod et al. explored the impact of these captions when viewed by people with vision impairments, finding that they are not sufficiently accurate [15]. They also evaluated ways of expressing the uncertainty in the caption model to engender skepticism when viewed as alt text. Automated approaches are popular because they are fast and cheap, allowing platforms to deploy them at scale to make large swatches of the web accessible for screen readers. However, they are often less descriptive compared to human-written alternative text on websites that prioritize accessibility and we consider how to combine strengths of both types of methods.

Crowdsourcing and Reusing Alternative Text

Human-in-the-loop systems can generate accurate alt text of images by soliciting descriptions from sighted crowd work-
While Twitter does allow users to add alt text to their images, ALT-Server to collect alt text from image links, and stores alternative text for later use similar to ALT-Server. However, these methods have trade-offs between description quality and frequency of applicability across a wide range of content. We build on prior work by integrating a broad set of automatic, retrieval, and crowdsourcing-based methods tailored to make a high percentage of Twitter images accessible.

### Social Media Accessibility

The content posted on social media platforms has become more visual over time as they have prioritized images, videos, and animations (GIFs). Morris et al. estimated that 28.4% of tweets on Twitter contained some multimedia in 2015 [20]. While Twitter does allow users to add alt text to their images, Gleason et al. has found that only 0.1% of images contained alt text, as the feature is not on by default [8]. Tweets do contain additional post text, but this is not the same as describing the contents of an image, and Morris et al. found that only 11.2% of post text would also be a good image description. Facebook has addressed the issue of alt text at social media scale by deploying object recognition software to efficiently create a large quantity of images descriptions [27], but they often do not contain enough detail to fully meet the needs of people with vision impairments. For example, Facebook describes the third picture in Figure 1 (Scene Description) as “1 person”.

Because social media is uploaded by end-users and not website developers, the images are unlikely to be shared elsewhere on the web, and if so, may not be indexed quick enough for Caption Crawler to find. Additionally, few images on platforms like Twitter are links to external websites, except for image previews of external links (which the user cannot choose). Specialized types of images, such as screenshots or memes, are more common on social media [20]. Gleason et al. designed a method for creating accessible memes by matching images to crowdsourced descriptions and extracting unique text from the image using OCR [9]. Another method recognizes the facial expression in memes to convey its emotional tone [23].

When describing end-user uploaded images, existing methods have trade-offs between description quality and frequency of applicability across a wide range of content. We build on prior work by integrating a broad set of automatic, retrieval, and crowdsourcing-based methods tailored to make a high percentage of Twitter images accessible.

### Requesting Alternative Text

When the user loads the Twitter.com web interface, the browser extension observes new images loaded on the page. When an image associated with a tweet is loaded on the user’s timeline, the extension extracts the image URL, any existing alt text, and context of the tweet (e.g., tweet ID, user ID, tweet text). If the tweet contains a preview image and link to an external website, the extension also records the linked URL. The extension automatically requests alt text from the server using this data, without user input.

### Obtaining Alternative Text

The Twitter A11y server receives requests for alt text that include the image URL and tweet context and attempts to use up to six methods as applicable to fetch or generate the image descriptions. The methods are ordered to prioritize a quick response, with methods that take less time returning early if they result in alt text for the image. The monetary costs (i.e., API request or crowd payment) and total time taken to return a response if a method is successful is presented in Table 1. These are the costs incurred per Twitter A11y response in our study, but we expect social media platforms or third-party application developers could implement these methods cheaper and faster. If an image has already been requested by another user and alt text exists in the database, that alt text is returned to the user immediately. Because of this, requests for tweets from popular Twitter accounts will have a very quick response time. If a tweet image contains existing alternative text written by the poster, it will not be replaced by Twitter A11y, as the original alt text is likely the most suited for the image.

If no alt text is present for the image, then Twitter A11y tries the following methods in order: URL Following, Text Recognition, Tweet Matching, Scene Description, Caption Crawler, Crowdsourcing (Figure 2). When a method returns alt text and the alt text satisfies the threshold conditions, the progression...
TEXT-BASED IMAGES. Many tweets feature images of text (e.g., text that exceeds the Twitter character limit, text messages, screenshots of tweets). We determine if an image tweet depicts primarily text (Figure 1, Text Recognition) using whole-image labeling via the Google Cloud Vision API [10]. If the confidence of the “text” label exceeds 0.8, we record the resulting Optical Character Recognition result (also via Google Cloud Vision API) as the alt text. We selected this threshold empirically to achieve high-precision (lower recall) such that users are unlikely to receive inaccurate alt text.

SCREENSHOTS OF TWEETS. Twitter provides a way for users to share other tweets by using the retweet feature, but users often share screenshots of tweets instead. This could be to preserve a tweet in case the author later deletes it or prevent harassment by sharing a tweet without notifying the original author. If the text recognition results from the previous step include a Twitter username beginning with the “@” symbol, we use the Twitter API to search for tweets by that user containing the first 10 words in the text recognition results. If a matching tweet is found, we describe that it is a screenshot of a tweet and return the tweet text directly, which can be more accurate than text recognition results.

REVERSE IMAGE SEARCH. Some images shared on social media are copied from elsewhere on the web, where alt text might be present. We re-implemented the Caption Crawler [11] project to source alt text from other websites with the same image. This method utilizes the Bing Image Search API [18] to perform a reverse image search. Once we have a list of locations where the same image appears around the web, we crawl up to 25 webpages to find the image and any alt text it contains. If multiple webpages have alt text, we return the longest one, as evaluation of the Caption Crawler project found length to be a good heuristic for image caption quality.

AUTOMATIC IMAGE CAPTIONING. Other social media platforms, such as Facebook, have experimented with providing image captions automatically generated by object recognition or scene description algorithms [27]. These are useful because they can be easily scaled to many images, but can lead blind users astray if the generated caption does not accurately or fully describe the image [15]. Twitter A11y will attempt to use this method if the previous ones were not applicable or did not result in alt text. It uses the Microsoft Cognitive Services Vision API to generate an image caption, and chooses the resulting caption with the highest score. If no caption exceeds a threshold of 0.7, determined empirically, then it is ignored.

CROWDSOURCE ALL REMAINING IMAGES. For remaining images not handled by the prior, rather quick methods, we post a crowd task on Amazon Mechanical Turk. This ensures that all requested images will receive some alt text from Twitter A11y. The task asks workers to generate image descriptions using the guidelines informed by Salisbury et al. [25]. The task time was originally estimated by the authors to take ~60 seconds and crowd workers were paid $0.17 per image ($10 per hour). After evaluation with some crowd workers, we found the median task to take 108 seconds (mean = 122s), so the task reward was increased to $0.30 per image. As the worker may take some time to write the description (~2-5 minutes), the browser extension displays “Waiting for crowd worker” until the written description is ready. Crowdsourcing carries the benefit that humans may be able to best describe characteristics such as humor that automatic methods miss.

DISPLAYING ALTERNATIVE TEXT
From the server, the extension receives either the existing alt text in the database, newly generated alt text from one of the above methods, or the status of an uncompleted crowdsourced description. The browser extension then dynamically inserts it into the alt text tag for the image. The user’s preferred screen reader can then read the newly generated alt text for an image when it focuses on the tweet, just as it would if the alt text had been there by default. While automatically-generated alt text appears soon after viewing (~2-10 seconds), crowd-generated alt text takes longer (on the scale of minutes) such that the user could view the text upon re-visiting the tweet (e.g., by scrolling back in their timeline, or revisiting a user’s page where the tweet appeared).

STATIC ANALYSIS OF BLIND USERS’ TIMELINES
To gather a large number of users who may use a screen reader to access Twitter, we examined the Twitter accounts following the National Federation for the Blind (@NFB_Voice) and the American Federation for the Blind (@AFB1921). We selected 50 users from this list who self-described themselves as blind or visually impaired in the profile description. It’s possible that these users do not use a screen reader to access Twitter,
<table>
<thead>
<tr>
<th>Strategy</th>
<th>Image</th>
<th>High-Quality Alt Text Rating</th>
<th>Low-Quality Alt Text Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td><img src="image1.png" alt="Image" /></td>
<td>Overhead map of the UK made up of people standing and the words During NEHW another 1,400 people across the UK will be diagnosed with advanced age-related macular degeneration</td>
<td>Palestinian protester</td>
</tr>
<tr>
<td>URL Following</td>
<td><img src="image2.png" alt="Image" /></td>
<td>A guide dog with a harness sits on the ground.</td>
<td>A point &amp; click adventure game about the fun, alienation, stupidity and agony of being a teen.</td>
</tr>
<tr>
<td>Text Recognition</td>
<td><img src="image3.png" alt="Image" /></td>
<td>UNDERSTAND THAT NOT EVERYTHING IS MEANT TO BE UNDERSTOOD. LIVE, LET GO, AND DON’T WORRY ABOUT WHAT YOU CAN’T CHANGE</td>
<td>Cittle oullorly Bllesengo l (lowserarch c Croaon)</td>
</tr>
<tr>
<td>Caption Crawler</td>
<td><img src="image4.png" alt="Image" /></td>
<td>Google home device pictured next to packaging box for size perspective</td>
<td>How to Make Special Video Effects</td>
</tr>
<tr>
<td>Scene Description</td>
<td><img src="image5.png" alt="Image" /></td>
<td>a group of people posing for a photo</td>
<td>a teddy bear sitting on top of a grass covered field</td>
</tr>
<tr>
<td>Crowdsourcing</td>
<td><img src="image6.png" alt="Image" /></td>
<td>A row of large, white Cannon professional video camera lenses are sitting on a perforated surface like a vent panel or an appliance.</td>
<td>Two faces hidden in the beautiful painting</td>
</tr>
</tbody>
</table>
and therefore they may not notice the presence or absence of alternative text. However, we are making the assumption that a large majority of these accounts that self-identify as blind or visually impaired do care about the presence of alt text on Tweets. The “Home Timeline” is the feed of tweets that a user reads when they log in to Twitter. We simulated a version of this timeline by collecting all of the tweets posted or retweeted by accounts each user followed over a 24 hour period. We then placed them in chronological order. All 50 users had at least 1,000 tweets in this simulated timeline.

These simulated timelines have some differences from those that would be experienced by users. First, Twitter does not always include all tweets in chronological order, choosing to place popular tweets first in the timeline. The Twitter algorithm may also hide replies to tweets that are not relevant, or include tweets liked by accounts the user follows. The actual timeline may also include ads. Finally, we were unable to collect tweets that had been deleted or were posted by protected (non-public) accounts, which the user may be able to see.

### Accessibility of Timelines

Accounts followed by blind users tend to include more alternative text in their tweets than Twitter as a whole [8]. For the 50 accounts, we randomly sampled 50 tweets from each user that either contained images or links to external website. From this 2,500 tweet sample, after examining the links, 1,041 tweets remained with either an image or valid link preview for a total of 1,198 images (a tweet can contain up to 4 images).

Of these 1,198 images, 62 contained alternative text from the tweet poster, and 29 link previews had alternative text from the linked website, meaning 7.6% was already accessible. We then evaluated each method except crowdsourcing on all of the images, and calculated the ability of each to add alt text for the same image, the Covered column does not sum to 100%.

<table>
<thead>
<tr>
<th>Alt Text Method</th>
<th>Covered N</th>
<th>Covered %</th>
<th>Unique N</th>
<th>Unique %</th>
<th>Selected N</th>
<th>Selected %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Alt</td>
<td>91</td>
<td>N/A</td>
<td>12</td>
<td>N/A</td>
<td>91</td>
<td>N/A</td>
</tr>
<tr>
<td>Scene Description</td>
<td>746</td>
<td>67.3</td>
<td>465</td>
<td>42.0</td>
<td>547</td>
<td>49.4</td>
</tr>
<tr>
<td>Text Recognition</td>
<td>216</td>
<td>19.5</td>
<td>60</td>
<td>5.4</td>
<td>199</td>
<td>18.0</td>
</tr>
<tr>
<td>Caption Crawler</td>
<td>213</td>
<td>19.2</td>
<td>70</td>
<td>6.3</td>
<td>70</td>
<td>6.3</td>
</tr>
<tr>
<td>URL Following</td>
<td>57</td>
<td>5.2</td>
<td>8</td>
<td>0.7</td>
<td>33</td>
<td>3.0</td>
</tr>
<tr>
<td>Tweet Matching</td>
<td>0</td>
<td>0.0</td>
<td>0</td>
<td>0.0</td>
<td>0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Crowdsourcing: 1107 100 258 23.3 258 23.3

Table 3. Evaluation of alt text methods in a sample of 1,198 images from blind users’ timelines. Covered indicates how many images in the sample the strategy could provide alt text for, and Selected indicates if that method was chosen according to Twitter A11y’s method priority order. Because different methods can provide alt text for the same image, the Covered column does not sum to 100%.

When evaluating methods we considered three metrics in addition to quality:

- **Image Coverage**: How many images did this method produce alt text for in the sample?
- **Method Uniqueness**: For how many images in the sample was this the only method that produced alt text?
- **Selected**: Using the Twitter A11y method priority as defined in Figure 2, how often was this method’s alt text chosen?

Using these metrics, we can change Twitter A11y’s method priorities to optimize for different aspects of the user experience. We can see that in all metrics, Scene Description is providing a bulk of the alt text, meaning that if optimizing for speed then it should be first in the method priority. Twitter A11y only used 70 of the images provided by Caption Crawler (because it was last in the order before crowdsourcing), while it was able to source alt text for up to 216. As automatic descriptions may have lower quality than human-written descriptions from Caption Crawler, perhaps it should be the “last resort” automatic method instead of Caption Crawler. We examine the quality of captions returned by each method next.

### Description Quality

Two members of the research team independently rated a random subset of the data collected, consisting of 50 images and alt text for each method (except URL Following and tweet matching as they did not have enough examples). We include sample descriptions for each method in Table 2. As in Gleason et al. [8], we utilized a four-point rating scale based on prior work [20, 25]. The scale ranged from “Irrelevant to image (0)” to “Great: almost everything described (3)” (available in Supplemental Material). To estimate agreement, we computed Cohen’s Kappa = 0.61, indicating substantial agreement [16]. The two raters then met and discussed each instance where their ratings differed until they reached agreement.

In Table 4 and Figure 3 these ratings are broken down by method. Crowdsourcing provided the highest “Great” quality descriptions as they were human-written with specific instructions for writing high-quality alt text. Text recognition has the highest “Great” percentage for an automatic method, comparable to alt text provided by the original poster and the only “Good” quality automatic captions from scene description. Surprisingly, Caption Crawler and URL following provided mostly “Irrelevant” or “Somewhat Relevant” alt text, even though they should source human written descriptions. Upon examination of the low quality descriptions, we found that this was typically because website authors were using an image filename or article title as the alt text for the image, instead of properly describing the visuals in the image itself (Table 2).

<table>
<thead>
<tr>
<th>Method (N)</th>
<th>Irrelevant</th>
<th>Relevant</th>
<th>Good</th>
<th>Great</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Alt (48)</td>
<td>4.2%</td>
<td>20.8%</td>
<td>25.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Scene Description (49)</td>
<td>8.2%</td>
<td>24.5%</td>
<td>53.1%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Text Recognition (49)</td>
<td>6.1%</td>
<td>16.3%</td>
<td>32.7%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Caption Crawler (38)</td>
<td>26.3%</td>
<td>42.1%</td>
<td>23.7%</td>
<td>7.9%</td>
</tr>
<tr>
<td>URL Following (26)</td>
<td>26.9%</td>
<td>46.2%</td>
<td>7.7%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Crowdsourcing (48)</td>
<td>10.4%</td>
<td>8.3%</td>
<td>18.8%</td>
<td>62.5%</td>
</tr>
</tbody>
</table>

Table 4. Two members of the research team rated a subset of the entire sample to estimate the quality of captions returned by each method.
13 participants who use a screen reader to access Twitter to with an average age of 33.5. Six participants were female and (P3, P9, P13) did not complete the data collection or final recruitment on Twitter. Participant ages ranged from 19 to 54, and seven were male. All participants accessed Twitter using a screen reader, although many said they used additional applications to access Twitter either on their computers or on their smartphones. These are detailed in Table 5.

Participants were compensated $10 for each interview, and $6 per 10 minutes of using the system, up to a maximum amount of $62 for the entire study.

Pre-study Interview
Before using Twitter A11y, we asked participants about their impressions of accessibility on Twitter, and what problems they saw with the social media platform and content on that platform. The specific questions can be found in the Appendix.

Regarding Twitter’s accessibility as a whole, ten participants said they found the platform itself mostly accessible. A few participants (3) complained about using a screen reader on the Twitter website, but stated the mobile applications were satisfactory. However, most participants (10) still choose to use third-party clients (e.g., Twitterific, TWBlue) either because they supported screen readers more effectively or because they did not change layouts often. TWBlue was especially popular, as it is an open source application designed with screen reader accessibility in mind. Participants lauded its support of global keyboard commands, meaning they did not have to switch applications to use it. They also liked that it included a function to request the text in an image using OCR.

The most common accessibility issue mentioned by every participant was media accessibility. They noted that most images did not include alt text, even though they likely followed accounts that added alt text more often compared to the sighted population. Twelve participants said they could sometimes guess at the content of some images depending on the text content of the tweet, but they commonly said this worked for less than half of images.

To understand how participants perceived the scope of the lack of alt text on image, we asked them to estimate what percent of their feed contained images and what percent of those images were accessible. On average, they estimated 50.5% of their feed contained images (max = 70%, min = 10%) and believed 12.1% of the images they encountered contained alt text (max = 30%, min = 1%). Eight of the participants mentioned that the percentage of images that people post affect their decision to follow an account, with some stating they would not follow an inaccessible account, some stating they would unfollow someone who did not add alt text after they requested it, and one stating they blocked those accounts to keep inaccessible images out of their feed.

Participants also complained that there was no mechanism to make GIFs (short animations), which are common on Twitter, accessible by adding alt text. When asked about video accessibility, responses were mixed, and many said they found videos consisting of mostly dialogue already reasonably accessible. Three participants suggested that videos on Twitter should support the addition of audio descriptions as a secondary audio track or as timestamped text content.

Some users circumvented the issues with image accessibility by utilizing other applications. Participants stated they used Microsoft Seeing AI on their iPhones to receive a textual description of an image (similar to Twitter A11y’s scene descriptions) or to read text in an image, if it was present. P12 noted that he could tell if an image contained text using his
Table 5. Demographics of participants who participated in the online study including age, gender, years on Twitter, level of vision, screen reader, methods of accessing Twitter, and other social networks used. Note that P3, P9, and P13 did not complete the data collection and post-study interview.

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>Years on Twitter</th>
<th>Level of vision</th>
<th>Screen reader</th>
<th>Twitter Applications</th>
<th>Other Social Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>19</td>
<td>M</td>
<td>6</td>
<td>Blind since birth</td>
<td>NVDA</td>
<td>TWBlue</td>
<td>Facebook</td>
</tr>
<tr>
<td>P2</td>
<td>39</td>
<td>F</td>
<td>11</td>
<td>Blind since birth</td>
<td>Voiceover</td>
<td>Twitterific</td>
<td>Facebook</td>
</tr>
<tr>
<td>P3</td>
<td>25</td>
<td>F</td>
<td>5</td>
<td>Blind since birth</td>
<td>NVDA, JAWS</td>
<td>TWBlue, Twitter (mobile site), Twitter for iOS</td>
<td>Facebook</td>
</tr>
<tr>
<td>P4</td>
<td>19</td>
<td>M</td>
<td>7</td>
<td>Blind since birth</td>
<td>NVDA, Voiceover, Talkback</td>
<td>TWBlue, Twitterific, Twitter for iOS, Twitter for Android</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>P5</td>
<td>32</td>
<td>M</td>
<td>12</td>
<td>Blind since birth</td>
<td>NVDA, JAWS</td>
<td>TWBlue</td>
<td>LinkedIn, Facebook, Instagram</td>
</tr>
<tr>
<td>P6</td>
<td>41</td>
<td>M</td>
<td>12</td>
<td>Blind since age 21</td>
<td>VoiceOver, JAWS, NVDA, Navigator</td>
<td>twitter.com, Twitterific, Twitter for iOS</td>
<td>LinkedIn, Facebook</td>
</tr>
<tr>
<td>P7</td>
<td>47</td>
<td>M</td>
<td>12</td>
<td>Blind since birth</td>
<td>JAWS, NVDA, VoiceOver</td>
<td>TWBlue, Twitterific, Twitter for iOS</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>P8</td>
<td>44</td>
<td>F</td>
<td>8</td>
<td>Blind since age 28</td>
<td>JAWS, NVDA, VoiceOver, Talkback</td>
<td>TWBlue, twitter.com, Twitter for iOS</td>
<td>Facebook, Instagram</td>
</tr>
<tr>
<td>P9</td>
<td>41</td>
<td>F</td>
<td>8</td>
<td>Blind since birth</td>
<td>VoiceOver</td>
<td>Twitter for iOS</td>
<td>None</td>
</tr>
<tr>
<td>P10</td>
<td>29</td>
<td>F</td>
<td>7</td>
<td>Blind since age 1</td>
<td>VoiceOver, JAWS</td>
<td>twitter.com, Twitter for iOS</td>
<td>Facebook</td>
</tr>
<tr>
<td>P11</td>
<td>22</td>
<td>F</td>
<td>8</td>
<td>Blind since birth</td>
<td>NVDA, VoiceOver</td>
<td>TWBlue, Twitter for iOS</td>
<td>Facebook</td>
</tr>
<tr>
<td>P12</td>
<td>23</td>
<td>M</td>
<td>2</td>
<td>Peripheral vision, no central vision since age 13</td>
<td>VoiceOver</td>
<td>Twitterific</td>
<td>Reddit, Youtube</td>
</tr>
<tr>
<td>P13</td>
<td>54</td>
<td>M</td>
<td>11</td>
<td>Totally blind since age 1.5</td>
<td>VoiceOver, NVDA</td>
<td>Twitterific</td>
<td>Facebook</td>
</tr>
</tbody>
</table>

Participants who used other social networks stated that the lack of image accessibility was common to Facebook, LinkedIn, Reddit, and Instagram. Most participants who used Facebook mentioned their automatic image alt text, stating that they wished they were more descriptive, but it was better than no alt text at all. P10 stated that she liked Facebook’s automatic captions simply because it stated “This image may contain: text”, so she knew to send the image to an OCR application.

Finally, we asked participants what they would do to make Twitter more accessible. Eight participants wished Twitter would make the alt text feature more prominent, ensuring all users are aware of the feature and how to use it. Three even suggested the feature should be mandatory, while 4 others wanted Twitter to fill in empty alt text with automatic descriptions similar to Facebook. Users who utilized third party applications stated Twitter should better support them, especially by ensuring all features (especially muting, blocking, and other reporting tools) were available via the API.

Twitter A11y Usage

After the interview, participants were directed to install Twitter A11y on their computers, and asked to use it for about 10 minutes a day over 7 days. Whenever they accessed Twitter, the browser extension logged every tweet containing images in their feed, and logged the alt text response added.

Session Length and Content

Participants used Twitter A11y for a total of 2,198 minutes over 145 sessions (mean = 15.2 minutes per session). During this time, they saw a total of 3,615 unique images (mean = 301.3). Of these, 86 already contained alternative text, and Twitter A11y provided descriptions for an additional 3,505 images. The breakdown of this by method and participant can be seen in Figure 4. The only time Twitter A11y did not provide any alternative text for an image was due to a technical error interrupting the request.

Post-study Interview

After 7-9 days had concluded, we asked the participants about their experiences with Twitter A11y. Overall, almost every participant stated that they enjoyed using the system and found that many images were more accessible with the alt text provided. P7 was the exception, as he stated he did not see much additional alt text on the images he viewed.

We asked each participant to rate the six methods used by Twitter A11y to generate alt text on a Likert item scale from Not useful at all (1) to Extremely useful (5) (see Table 6). When using Twitter A11y, the participants could tell which method generated a result, as each alt text was preceded by “From [method name]”. Some participants stated they did not see any examples of a particular method, in which case they were unable to give an answer. We also asked participants to order the methods they preferred from best to worst. Because the majority of participants only felt comfortable ranking scene description, text recognition, and crowdsourcing, we only show the mean ranks for those. We do not report statistical testing for these responses, as not all participants were able to rate or rank every method, resulting in a small sample size.
When assessing the usefulness of a description method as when assessing the usefulness of a description method for a particular tweet image, some participants reported the value of the participants’ perceived value of the method itself. Future work could attempt to separate these two factors. For example, when assessing the usefulness of a description method for a particular tweet image, some participants reported the value depended on both the content of a description along with the context of the surrounding tweet. As P1 describes:

Sentence: [Scene description] is useful when the tweet itself gives some context so if a tweet talks about a protest, and [scene description] says “a person holding a sign”, then that makes sense because I understand that they’re protesting but it doesn’t give a ton of detail.– P1

When assessing the usefulness of a description method as a whole, participants also considered the usefulness of the method for their particular timeline. As P4 reports:

Sentence: [Text recognition] is one of the most useful because a lot of my timeline is tech and geek stuff which often has screenshots. Text recognition is key to understanding what’s going on. – P4

Inaccurate descriptions could lead people with vision impairments to believe a tweet image contained something that was not actually present. We asked participants if they ever felt like they did not trust a caption because it seemed inaccurate. Several participants stated they read 1-2 captions from scene description method that did not make sense with the context of the rest of the tweet, but otherwise everyone said they assumed the tweets were accurate. This implies that future versions of Twitter A11y should integrate cautionary text explored by MacLeod et al. [15] to encourage distrust in automatic captions when uncertainty is high.

In general, participants felt Twitter A11y could be most improved by speeding up the responses from crowdsourcing, as they did not want to wait minutes for a worker to describe the image. Participants also wanted flexibility to choose the method by which an alt text is generated, possibly through keyboard commands or a menu. Additionally, they wanted to be able to use multiple methods together, specifically scene description and text recognition, to get a sense of the image and recognize any specific text. Finally, participants were eager to see Twitter A11y integrated into their preferred clients, as they find them more usable than the Twitter web interface.

**DISCUSSION**

The participants in our user study expressed that Twitter A11y offered an impressive level of accessibility compared to what they typically find on social media platforms. This mirrors our findings from the analysis of static sample of tweets in blind users’ timelines, which increased the presence of alt text from 7.6% to 78.5%, with an estimated 57.5% of descriptions being rated “Good” or “Great”. We integrate our findings from these analysis to discuss the alt text generation methods holistically.

Participants preferred the text recognition and automatic captioning methods because they were quick (~2.5s) and often descriptive (~67-77% “Good” to “Great”). While most participants were familiar with these methods from other applications, they expressed that having the alt text automatically attached to images made their experience much more accessible. In our static analysis, we see the crowdsourcing provides the highest percentage of “Great” alt text (62.5%), but offers the lowest value (only 2-3 “Good” or “Great” captions per dollar) due to the expense of paying human annotators. Participants also perceived crowdsourced descriptions as accurate, but too slow (~2 minutes) to wait for when browsing social media sites. The tweet matching, URL Following, and Caption Crawler methods were highly rated by the participants who encountered them, but results with the methods were too rare for all participants to form an opinion. However, we only used Caption Crawler when no other automatic methods produced a result (6.3% of sample), and the coverage results in the static analysis indicate it had results for many more images (19.2% of sample) that Twitter A11y did not use, suggesting that the priority of methods could be re-examined.

For social media platform designers and application developers seeking to add automatically generated alt text, we would recommend using methods that produce cheap, high-quality results first. This would include text recognition, followed by scene description methods. Caution should be used when integrating the latter, as prior research has shown that inaccuracies are not easily noticed by people with vision impairments [15, 25]. Other methods do produce additional alt text, including URL following and Caption Crawler, but the low-quality results indicate they should be a low priority. Crowdsourcing is clearly the solution that produces the highest-quality alt text, but asking crowd workers to label images is likely prohibitively expensive at scale. Instead, designers and developers should explore if they can design features to support friends and other volunteers in adding alt text [4, 5].

Several participants indicated that other social media platforms include a higher frequency of inaccessible images, including Facebook and Instagram. We designed Twitter A11y specifically for evaluation on the Twitter website, but there is strong indication that this tool would be useful on other websites. Participants were unanimous in their belief that Twitter A11y would work equally well if deployed on other social media platforms they used, and the methods that provided high coverage of images and high-quality captions are readily applicable to other platforms. The only method that could not be easily re-engineered for other platforms is tweet matching, which was not used in the static evaluation as screenshots of tweets are a rare image category.

Two participants in our user study raised the importance of distinguishing accessibility and accommodation. They viewed

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Usefulness</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caption Crawler</td>
<td>4.8</td>
<td>N/A</td>
</tr>
<tr>
<td>Tweet Matching</td>
<td>4.5</td>
<td>N/A</td>
</tr>
<tr>
<td>URL Following</td>
<td>4.3</td>
<td>N/A</td>
</tr>
<tr>
<td>Text Recognition</td>
<td>4.3</td>
<td>1.3</td>
</tr>
<tr>
<td>Automatic</td>
<td>3.7</td>
<td>3.1</td>
</tr>
<tr>
<td>Crowdsourcing</td>
<td>3.6</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Table 6. Participant ratings of methods on a scale from Not at all useful (1) to Extremely useful (5). Participant mean rankings for the three most common methods (with 1 being their first choice).

Note that this self-reported metric of “usefulness” does not distinguish between the quality of the alt-text received and the participants’ perceived value of the method itself. Future work could attempt to separate these two factors. For example, when assessing the usefulness of a description method for a particular tweet image, some participants reported the value depended on both the content of a description along with the context of the surrounding tweet. As P1 describes:

Sentence: [Scene description] is useful when the tweet itself gives some context so if a tweet talks about a protest, and [scene description] says “a person holding a sign”, then that makes sense because I understand that they’re protesting but it doesn’t give a ton of detail.– P1

When assessing the usefulness of a description method as a whole, participants also considered the usefulness of the method for their particular timeline. As P4 reports:

Sentence: [Text recognition] is one of the most useful because a lot of my timeline is tech and geek stuff which often has screenshots. Text recognition is key to understanding what’s going on. – P4

Inaccurate descriptions could lead people with vision impairments to believe a tweet image contained something that was not actually present. We asked participants if they ever felt like they did not trust a caption because it seemed inaccurate. Several participants stated they read 1-2 captions from scene description method that did not make sense with the context of the rest of the tweet, but otherwise everyone said they assumed the tweets were accurate. This implies that future versions of Twitter A11y should integrate cautionary text explored by MacLeod et al. [15] to encourage distrust in automatic captions when uncertainty is high.

In general, participants felt Twitter A11y could be most improved by speeding up the responses from crowdsourcing, as they did not want to wait minutes for a worker to describe the image. Participants also wanted flexibility to choose the method by which an alt text is generated, possibly through keyboard commands or a menu. Additionally, they wanted to be able to use multiple methods together, specifically scene description and text recognition, to get a sense of the image and recognize any specific text. Finally, participants were eager to see Twitter A11y integrated into their preferred clients, as they find them more usable than the Twitter web interface.

**DISCUSSION**

The participants in our user study expressed that Twitter A11y offered an impressive level of accessibility compared to what they typically find on social media platforms. This mirrors our findings from the analysis of static sample of tweets in blind users’ timelines, which increased the presence of alt text from 7.6% to 78.5%, with an estimated 57.5% of descriptions being rated “Good” or “Great”. We integrate our findings from these analysis to discuss the alt text generation methods holistically.

Participants preferred the text recognition and automatic captioning methods because they were quick (~2.5s) and often descriptive (~67-77% “Good” to “Great”). While most participants were familiar with these methods from other applications, they expressed that having the alt text automatically attached to images made their experience much more accessible. In our static analysis, we see the crowdsourcing provides the highest percentage of “Great” alt text (62.5%), but offers the lowest value (only 2-3 “Good” or “Great” captions per dollar) due to the expense of paying human annotators. Participants also perceived crowdsourced descriptions as accurate, but too slow (~2 minutes) to wait for when browsing social media sites. The tweet matching, URL Following, and Caption Crawler methods were highly rated by the participants who encountered them, but results with the methods were too rare for all participants to form an opinion. However, we only used Caption Crawler when no other automatic methods produced a result (6.3% of sample), and the coverage results in the static analysis indicate it had results for many more images (19.2% of sample) that Twitter A11y did not use, suggesting that the priority of methods could be re-examined.

For social media platform designers and application developers seeking to add automatically generated alt text, we would recommend using methods that produce cheap, high-quality results first. This would include text recognition, followed by scene description methods. Caution should be used when integrating the latter, as prior research has shown that inaccuracies are not easily noticed by people with vision impairments [15, 25]. Other methods do produce additional alt text, including URL following and Caption Crawler, but the low-quality results indicate they should be a low priority. Crowdsourcing is clearly the solution that produces the highest-quality alt text, but asking crowd workers to label images is likely prohibitively expensive at scale. Instead, designers and developers should explore if they can design features to support friends and other volunteer in adding alt text [4, 5].

Several participants indicated that other social media platforms include a higher frequency of inaccessible images, including Facebook and Instagram. We designed Twitter A11y specifically for evaluation on the Twitter website, but there is strong indication that this tool would be useful on other websites. Participants were unanimous in their belief that Twitter A11y would work equally well if deployed on other social media platforms they used, and the methods that provided high coverage of images and high-quality captions are readily applicable to other platforms. The only method that could not be easily re-engineered for other platforms is tweet matching, which was not used in the static evaluation as screenshots of tweets are a rare image category.

Two participants in our user study raised the importance of distinguishing accessibility and accommodation. They viewed
Twitter A11y’s efforts as important to provide reasonable accommodations for images that were not made accessible from the start. However, they were not willing to use this tool unless it also made an effort to increase alt text provided by end-users who uploaded photos. The image posters have important contextual knowledge, and even the best crowd worker will not fully understand their intent when posting the image or all important details (i.e., names). The participants suggested that Twitter A11y automatically notify the image poster that a blind user found their post inaccessible, and provide instructions on how to add image descriptions on Twitter. We agree with the participants that social media platforms should consider additional accessibility features and user education that could improve accessibility, not just rely on accommodations such as scene description methods. Some recommendations specific to Twitter are to increase enable image descriptions by default, train users on what comprises good alternative text, and give users feedback on the alternative text they write [8].

As members of the research team (all sighted) tested Twitter A11y, we were surprised at how useful alt text could be even in conjunction with seeing the image, indicating the tool could provide value for sighted users. The image captions served as quick summaries of a scene, and provided additional context. Specifically URL following, scene description, and Caption Crawler often added the names of people and places or described events that were not easily discernible from the images (see Figure 1). Additionally, when an image contained alt text written by the original image poster, it served as an indication that they valued accessibility for people with vision impairments. In contrast, the constant lack of original alt text served as a reminder that the majority of images are inaccessible and why image descriptions are valuable.

Limitations & Future Work
The major limitation with our evaluation of Twitter A11y is the rather short week-long evaluation with small number of participants (10) completing the study. A longitudinal study with more participants would be necessary to understand any behavior change of Twitter users due to increased accessibility of images. Additionally, we asked participants to use the Twitter web interface, which was typically not their preferred client, so an evaluation of Twitter A11y that was more tightly integrated with their Twitter client of choice would likely yield results more representative of typical use. Finally, a major area for future work is validating that our rubric of alt text quality aligns with the expectations of blind users. While the rubric was constructed based on prior work with blind social media users, it has not been validated to ensure that the 4-point rating scale accurately captures different levels of “usefulness” that people with vision impairments might desire.

Our interviews with participants and evaluation of a tweet sample indicate other avenues for future work. First, participants raised the desire for an integration of multiple methods, such as scene description and text recognition. Additionally, Twitter A11y currently tries each method in a sequential order until an alt text result is found, but our static evaluation revealed overlap between some methods. If there was a clear approach to score the quality of image descriptions from multiple methods, Twitter A11y could ensure the best alt text is always returned. Gleason et al. [8] briefly suggested generating automatic feedback for users while they write image descriptions based on the post text and the image description. The inclusion of these language features and features from the image itself could be adapted to develop a ranking algorithm.

We only address image accessibility in Twitter A11y, but other forms of visual media, such as GIFs and videos, were reported by participants to be inaccessible on social media platforms. As GIFs are short looping animations, they straddle the line between a static image and a longer video. An exploration of the best way for Twitter A11y to make these accessible might explore the use of an alt text description versus an audio description that describes the action in the GIF.

It is unlikely that the methods we tested will be directly applicable to creating audio descriptions, so new avenues will need to be explored to address video inaccessibility. There is not yet a robust method for automatic description of actions in videos [21], but there is a dataset of audio descriptions for movies to encourage future research on generating video descriptions [24]. Additionally, the YouDescribe project [22] has demonstrated how dedicated volunteers can describe videos and share audio descriptions through browser extensions, meaning Twitter A11y could explore a crowd workflow for generating audio descriptions on social media networks.

CONCLUSION
The lack of accessible content on social media platforms is a major barrier for participation by people with disabilities. Our participants echoed this in their interviews, stating that it was their primary concern and they often had to find workarounds for images without alt text. Making the deluge of user-generated content accessible, at scale, seems challenging, but platforms such as Facebook are attempting this.

Twitter A11y represents an attempt to merge promising methods for finding or generating new alternative text into one tool that users can use on Twitter. We demonstrated Twitter A11y’s ability to take the content followed by a blind user from 7.6% to 78.5% with accessible images. Of these images, 57.5% of descriptions receive a “Good” or “Great” quality rating.

This tool represents a large leap in making content on these major platforms accessible, and we believe it could be easily modified, refined, and deployed on other social media platforms that include images with limited alternative text (Instagram, Reddit). We also encourage social media platforms to take note of the success of some of these methods, especially text recognition and automatic captioning, and integrate them into their platforms to improve accessibility for people with vision impairments.

ACKNOWLEDGMENTS
This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (#DGE-1745016). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
REFERENCES


[14] @JohnMu. 2018. Alt text is extremely helpful for Google Images – if you want your images to rank there. Even if you use lazy-loading, you know which image will be loaded, so get that information in there as early as possible & test what it renders as. (4 September 2018). https://twitter.com/JohnMu/status/1036901608880254976


[21] Jaclyn Packer, Katie Vizenor, and Joshua A. Miele. 2015. An Overview of Video Description: History,
APPENDIX

PRE-STUDY INTERVIEW QUESTIONS

1. Demographics information:
   (a) Age
   (b) Gender
   (c) Years using Twitter
   (d) Visual ability
   (e) Any other disabilities?
   (f) Years of vision disability
   (g) What screen readers do you use?
   (h) What other social networks do you use?

2. How accessible do you find Twitter?
3. What are major barriers for using the site?
4. What percent of images that you encounter on Twitter do you think have an image description?
5. Do you find it easy to understand tweets that include an image with no description?
6. Does accessibility of images affect which accounts you follow?
7. What about other forms of media, such as videos and GIFs; how accessible are those?
8. What changes would you make to Twitter to make it more accessible for people with vision impairments?
9. Do you use any other tools to make Twitter more accessible?

POST-STUDY INTERVIEW QUESTIONS

1. What was your experience using the tool? Did you enjoy it?
2. What percent of image do you think were accessible when you used the tool and browsed Twitter?
3. We used many methods to make Twitter images accessible. Which did you find the best? The worst?
4. Were you ever uneasy or felt like you didn’t trust a description?
5. Would you continue using a tool like this?
6. What could be most improved about this tool? What worked well?
7. Do you think this tool would translate well to other social networks you use?
8. What else should we focus on to make social networks more accessible?