EyeDescribe: Combining Eye Gaze and Speech to Automatically Create Accessible Touch Screen Artwork

Kyle Reinholt, Darren Guinness, Shaun K. Kane
University of Colorado Boulder
Boulder, United States
{kyle, darren.guinness, shaun.kane}@colorado.edu

ABSTRACT
Many images on the Web, including photographs and artistic images, feature spatial relationships between objects that are inaccessible to someone who is blind or visually impaired even when a text description is provided. While some tools exist to manually create accessible image descriptions, this work is time consuming and requires specialized tools. We introduce an approach that automatically creates spatially registered image labels based on how a sighted person naturally interacts with the image. Our system collects behavioral data from sighted viewers of an image, specifically eye gaze data and spoken descriptions, and uses them to generate a spatially indexed accessible image that can then be explored using an audio-based touch screen application. We describe our approach to assigning text labels to locations in an image based on eye gaze. We then report on two formative studies with blind users testing EyeDescribe. Our approach resulted in correct labels for all objects in our image set. Participants were able to better recall the location of objects when given both object labels and spatial locations. This approach provides a new method for creating accessible images with minimum required effort.

Author Keywords
Accessibility; eye gaze; touch screens; blindness; visual impairments; image captioning.

ACM Classification Keywords
Human-centered computing → Accessibility technologies

INTRODUCTION
Over the last two decades, the number of digital images on the Web has increased at an incredible pace. There are more than 4 billion total pages on the Web [24]. Approximately 1.8 billion images are added to social media pages each day [10,28], and this figure only accounts for a subset of the Web (Facebook, Twitter, Instagram, LinkedIn, etc.).

Many of the images on the Web represent items of cultural significance which are shared freely so that everyone can access them. These images include works of art from museum collections such as the Museum of Modern Art, The Smithsonian Institution, and The Metropolitan Museum of Art. The Metropolitan Museum of Art alone has catalogued up to 5,000 years' worth of artwork as digital images. Knowing about popular art is a part of participating in shared culture. Often, however, these images are not accessible to someone who is blind or visually impaired.

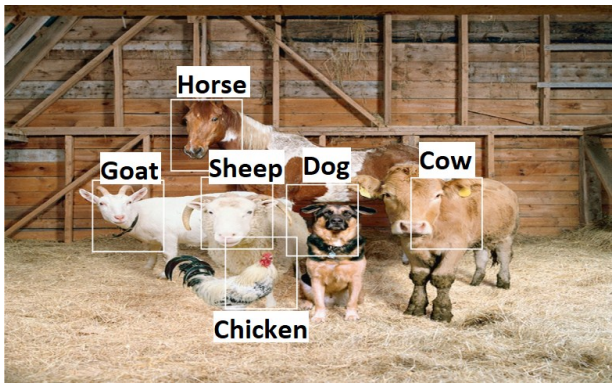

Figure 1. Using eye gaze and speech data, EyeDescribe identifies regions of interest in an image, which can then be presented interactively to blind and visually impaired end users via a talking touch screen application.
Image Copyright Rob MacInnis.

When a visually impaired person encounters an image on the Web, their experience is strongly affected by whether the image authors have done the work to include accessibility features. In particular, making an image accessible online involves creating an alternative text description (or "alt text") that describes the image [29]. However, even when text descriptions are provided, they do not always capture what a user might wish to know about the image.

In the case of artwork, text descriptions are often designed for the sighted audience, providing historical details rather than describing the image itself [30]. Some online art exhibits use audio description (AD) [33] to provide verbal descriptions of media to enhance a blind viewer's experience, but many online image sources do not put in the extra effort to do so [36]. Furthermore, most text or spoken descriptions lack spatial information describing the positions of elements within the image [20,21,28]. Thus, while a description of abstract art may describe the image elements present, most

ISS'19, November 10–13, 2019, Deajeon, Republic of Korea.

ACM ISBN 978-1-4503-6891-9/19/11...$15.00.
DOI: http://dx.doi.org/10.1145/10.1145/3343055.3359722

descriptions do not detail information about the layout of the image that is available to sighted people.

Our work is motivated by these three concerns: the lack of available image descriptions, the effort needed to create new image descriptions, and the lack of spatial information in descriptions. In an ideal world, image descriptions should be easy to create and should contain the same information that is available to a sighted viewer. While manually generating these types of descriptions requires effort, we were motivated by the belief that descriptive information about an image is implicitly created when a sighted person views an image. For example, tracking a viewer's eye gaze may identify important areas in an image, and capturing the viewer's discussion of an image may provide a usable description. This approach, which we call ***use-driven accessibility,*** offers the potential to create rich descriptions of images with minimal effort from the image creator.

In this paper, we explore the application of use-driven accessibility to artistic images. We developed a prototype system called EyeDescribe, which captures gaze and speech data as a sighted person views an image and automatically generates a rich, spatialized description of that image (Figure 1). We describe the iterative design, development, and testing of EyeDescribe with blind users. Our user studies explore the following research questions:

RQ1. How can we best use everyday gaze and speech data to generate spatial captions?

RQ2. How do spatial captions generated by EyeDescribe compare to traditional text descriptions?

RQ3. How do spatial captions affect a user's ability to reconstruct scenes in an image?

The contributions of this paper are: 1) introducing a use-driven accessibility approach to generate accessible images; 2) introducing EyeDescribe, an application to create image descriptions from eye gaze and speech data; 3) documenting two rounds of user studies in which blind participants evaluated image descriptions created using EyeDescribe.

RELATED WORK

Accessible Image Representations

A common way to represent spatial information is with a tactile graphic [9]. Most tactile graphics are produced via embossed Braille or tactile swell paper, although they may also be fabricated using 3D printers [5,37]. Tactile graphics provide a spatial overview of an image and may include Braille labels. More technically advanced tactile graphics may include interactive audio labels [4,31,35].

While tactile graphics provide rich information, they require expensive hardware to produce [18]. Designing tactile graphics also requires significant training and skill [37]. As tactile graphics are physical items that must be printed, they are not especially portable and require time and advance planning to produce.

Another approach for creating accessible image representations is through sonification, in which content in the image is represented through speech and non-speech audio. Sonified images can be made interactive by placing them on a touch screen [20]. While sonified images can provide both spatial and descriptive data, exploring large or complex images can become overwhelming [7]. This complexity can sometimes be mitigated by adding new gestures, such as in work by Rector et al. [30] that leveraged a proxemic spatial interface in which the user could move closer to or further from an image to hear different information about the image. In each of the previous examples, image labels were generated by humans. Our approach builds on prior work in spatialized and sonified images, but explores new ways to create spatial image captions.

Creating Image Captions

Even though many images on the Web are not accessible, the Web offers the largest collection of labeled images. Web page creators can add a description to their images by adding alternative text tags to their HTML document [40]. One notable limitation of Web-based image captions is that they typically include only a text string describing the image, which may overlook important spatial or visual characteristics of the image [28].

Even the simple task of describing Web images presents several barriers. First, as labeling an image is sometimes considered extra work (or even unnecessary work), many authors do not take the time to label their images. If an author chooses to add an image, he or she may not know how to write an effective caption [2].

Some tools exist to suggest image descriptions or even to generate descriptions for non-captioned images. WebInSight [3] used optical character recognition to automatically recognize text in Web images. Caption Crawler [15] is a browser plugin that identifies non-captioned images and searches the Web for other instances of that image that are captioned, bringing the caption from another site (with better accessibility) to the user's current site. In addition to these techniques, crowdsourcing can also be used to recruit human image labelers [19]. YouDescribe[1] provides a platform that allows users to create and request audio descriptions of online videos. Finally, recent advances in machine learning have led to automated systems such as Microsoft's CaptionBot, Facebook's Automated Alt-Text, and Google's Vision API, which can automatically generate an image description. These automated approaches are helpful, but can sometimes fail in frustrating ways [26,38]. Furthermore, these approaches mostly focus on listing the types of objects in an image, and do not currently provide information about the spatial layout of the image, nor do they provide the type of subjective information that may be useful when viewing artistic images. Our proposed work complements these

[1] https://youdescribe.org

approaches and introduces a new resource for creating image descriptions.

Eye Gaze Input

Most eye tracking technology uses infrared imaging to track movement of the user's gaze over time [8]. Eye trackers can detect the rapid, unconscious movements of the eyes (called saccades); however, it is the longer periods without movement (called fixations) that typically correspond to the user's conscious visual attention [14,32].

For more than a century, researchers in the social sciences have observed eye gaze as a way to better understand human cognition and behavior [16]. The first eye tracking device was used by Huey in 1908. In 1932, another, less intrusive, eye tracker was introduced and was used to collect data about how people viewed photographs [6]. For most of its history, eye gaze has been used as another form of behavioral data. For example, eye tracking has been used to detect tired drivers [12], track attention during educational activities [17], and for biometric authentication [23].

Much eye gaze research has focused on tracking a user's gaze as they observe an image, and some of this work has explored how eye gaze data can help provide information about the image itself. Eye gaze data has been used to crop images [34] and as auxiliary input to a computer vision-based object recognizer [22].

While most of this research has focused on using eye gaze data alone, Yun et al. studied the relationship between a participant's gaze points and their simultaneous description of an image [39], comparing the descriptive text to a set of previously annotated images [11]. This work examined both whether people fixate on objects as they describe them as well as whether they describe the objects that they fixate upon. This study showed that people fixated on what they described 87% of the time and described what they fixated on 95% of the time [39]. This research confirms that gaze and speech together provide a useful signal for both locating and identifying objects in an image. In this work, we leverage the signal generated by users who view an image, with the goal of creating accessible images with minimal effort from the labeler.

USE-DRIVEN ACCESSIBILITY

In most cases, creating image descriptions requires conscious effort from a human labeler. The labeler must examine the image and generate an appropriate description. This work often requires that the labeler predict the future use of the image. For example, a person who labels a family portrait must decide which information to include or exclude from the description, including both objects in the image as well as seemingly incidental details about the position of people in the image, clothing, facial expressions, and so on. Furthermore, as most current image description tools generate a plain text description, the labeler must decide whether to describe the locations of objects or the spatial relationships between them. For complex images, such as those found in collections of visual art, creating a comprehensive description may require extensive work from the labeler.

In this work, we consider how a user's undirected interaction with an image (i.e., simply viewing the image), can automatically provide information about the content of that image. As an example, imagine a technical presentation in which the presenter is discussing an on-screen diagram. As the presenter discusses part of the diagram, they will often look at that area, and may even point toward that area to guide the audience's attention. Combining these information sources may indicate both what objects are in the diagram as well as their location within the diagram. The approach follows a similar motivation to the ESP Game [1], which on the surface is a fun game to play, but implicitly generates labeled images for people who are blind or visually impaired.

We call this approach ***use-driven accessibility,*** as the image's accessible description arises from the sighted user's interaction with the image, rather than any conscious work to create an image description. Our goal in this paper is to explore whether a sighted user's recorded interactions with an image can be used to annotate that image.

Use-Driven Spatial Captioning

As a proof-of-concept of use driven accessibility, we explored the use of interaction data to create a spatially annotated image. We define **spatial captioning** as the creation of an image description that includes both text describing the image and metadata describing the location of objects referenced in the description. As noted previously, an image that contains both text descriptions and spatial labels can enhance a blind person's experience of that image [20,21,28].

In the general case of spatial captioning, we consider the ***input*** to be a two-dimensional image (and other relevant metadata), and the ***output*** to be a list of labeled spatial regions. Each spatial region includes a text description of objects within the region as well as spatial information such as a bounding box or center point.

EyeDescribe's Approach to Spatial Captioning

To create spatial captions for an image, EyeDescribe assumes access to the following data:

1) A two-dimensional image.
2) Eye gaze data from one or more sighted users ("gazers"). This data consists of a series of (x,y) locations and associated timestamps.
3) Text or speech data with a corresponding timestamp. This speech may correspond to a gazer's comments about the image or may reflect an audio description heard while gazing at the image.

The expected output of this approach is a series of spatial captions that indicate some spatial region in the image, along with a text label of that region. In this study, we have primarily focused on captions that describe the objects in the

image, as this information is useful to blind viewers. However, this approach could also be used to extract other sorts of descriptive terms. For example, spatial captioning of an abstract art painting might instead focus on descriptions of colors, textures, or even emotions felt by the gazer.

In the following sections, we describe our development of the EyeDescribe system and how it may be used to generate spatial captions from sighted users' data. Although there are many ways to evaluate the quality of image captions, we generally focus on whether this approach leads to correct labels at the correct locations.

Once spatial captions have been created, we must still solve the problem of how to present these captions and how a user may interact with a spatially captioned image. After documenting the creation of spatial captions using EyeDescribe, we then explore how this information may be presented non-visually to a blind or visually impaired person.

FEASIBILITY STUDY

To demonstrate that simultaneous gaze and speech data can be used to create spatial captions, we first conducted a feasibility study in which we collected eye gaze data from 10 sighted users as they each viewed 10 different images representing paintings and photographs.

While we believed that the combination of eye gaze data and speech data would be useful, we were unsure about how to best collect this data. To understand what conditions are necessary to create spatial captions, we explored three scenarios for collecting users' eye gaze data: gazing at an image in silence, gazing at an image while listening to a spoken description of that image, and gazing at an image while the gazer describes the image themselves.

Initial Data Collection

For this study, a set of 10 images was used, containing a mix of paintings and photographs. Images were chosen from a set of online art gallery collections and were intended to capture a range of image types. We originally intended to pair these images with text descriptions from their source galleries. However, we soon found that most of the descriptions that accompanied these images focused on the historical context of the image and did not describe the visual content of the image. To create a neutral set of image descriptions, we elicited a set of text descriptions from crowd workers on Mechanical Turk using the prompt:

> Please write a description for each of the 10 images below. This description should be somewhere between 4 and 10 sentences. Your text should describe all the important items in the image.

After collecting descriptions from 6 Mechanical Turk workers, we chose an image description randomly from the set of descriptions collected for each image. The average number of chosen sentences per image description was 6; sentences were 12 words long on average.

These descriptions were presented to study participants via audio. Each description was rendered to audio using the text-to-speech engine on MacOS.

Participants

Participants were recruited from a local university. The inclusion criteria were that the participant had normal (or corrected-to-normal) vision and that they were 18 years or older. We recruited 10 participants (3 female) with a mean age of 29 (range 23–37).

Apparatus

Gaze and speech data were collected using a MacBook Pro laptop running Windows 10. The laptop displayed images on a 13-inch screen (2500×1600 resolution). We used the Tobii 4C eye tracker to collect eye gaze data, which ran at 90 Hz.

in a comfortable position that they could maintain for at least 30 minutes. The participant then completed the eye gaze calibration process provided by the Tobii eye tracker software, which involves gazing at about 4 on-screen targets.

Following the calibration process, the participant completed three study activities, each of which involved gazing at the test images. The order of images was randomized for each participant, but the order of the three activities was kept the same for all participants. Between each image, the participant viewed a blank image to clear their palate and pressed a button to advance to the next image. Study sessions took approximately 30 minutes to complete.

Activity 1: Looking in Silence

During this activity, the participant gazed at each image in silence for 30 seconds.

Activity 2: Looking while Listening

During this activity, the participant gazed at the image while listening to the text-to-speech version of the image description.

Activity 3: Looking while Describing

During this activity, participants gazed at the image while describing the image in their own words.

Analysis

We recorded eye gaze data for all three activities. For Activity 3, we recorded the participant's speech and transcribed their spoken descriptions of the image using IBM Watson's Speech-to-Text[2] API, which provides the transcribed word as well as a timestamp. The files were manually corrected for recognition errors by the research team. By aligning the eye gaze timestamps and speech timestamps, we were able to identify the user's gaze location while speaking a particular word. We also collected timestamps from the text-to-speech recordings of the Mechanical Turk descriptions so that we could identify where a participant was looking when they heard a particular word.

[2] https://www.ibm.com/watson/services/speech-to-text

Collecting and Aggregating Gaze Data

We explored several methods for analyzing the collected gaze data:

Raw Gaze Analysis. The raw data stream from the eye tracker, including both conscious fixations and unconscious saccades. Each data point included the gaze *x*-coordinate, gaze *y*-coordinate, and a timestamp.

Fixation Analysis. Filtered data that includes only fixations. For each fixation, we recorded the gaze *x*-coordinate, gaze *y*-coordinate, timestamp, and duration. Fixations were calculated using the Dispersion Threshold Identification algorithm [32].

For raw gaze and fixation, we looked at *what percentage of gaze points* were within a specified bounding box (e.g., what percentage of gaze points were within the bounding box of "oranges" while hearing or saying "oranges").

Raw Kernel Density Estimation (Raw KDE). We calculated a kernel density estimation on the raw gaze points to return one (*x,y*) coordinate that corresponds to the densest location in the raw gaze data.

Fixation Kernel Density Estimation (Fixation KDE). We calculated a kernel density estimation over the collection of fixations to return one (*x,y*) coordinate that corresponds to the densest location among fixations.

For the KDE metrics, we calculate a single central point per utterance (i.e., what is the center of gaze when listening to or speaking "oranges"). We then measure *how many times* that central point is within the bounding box (e.g., how many times a participant's center of gaze was within the bounding box for "oranges" while hearing or saying "oranges").

Associating Gaze and Speech Data

Participants' speech utterances were segmented into separate sentences for Activity 2 using the sentences generated from Mechanical Turk descriptions. For Activity 3, only words were segmented because of sentence ambiguity in spoken language. Each word and sentence, and their associated timestamps, were stored in a database. This allowed us to calculate the densest gaze point at the time an utterance was heard or spoken using Kernel Density Estimation.

Because speech and gaze can be out of sync [27], we tested extending the time window in which we collected eye gaze data relative to the start and end of the word. we aggregated the data for each activity (across all participants), and adjusted the window by 50 milliseconds (up to 1 second on either side) until we found the highest Raw KDE accuracy. For Activity 2, the Raw KDE accuracy was highest with a time window starting at the beginning of a word up to 500 milliseconds after the word. For Activity 3, the Raw KDE accuracy was highest with a time window starting 1 second before a word up to the end of the word.

Identifying Objects in the Image

To analyze when the user is looking at a particular object, we generated ground truth data for the objects in the image. Two colleagues manually annotated objects and features in each of the test images. These colleagues were familiar with the general purpose of the research but did not see any data collected from actual users. These annotators used the web application Annotorious[3] to identify regions in the image and assign labels to them. These annotators chose their own labels for each region.

We chose the ground truth objects for each image by combining the data collected by Mechanical Turk crowd workers, study participants, and our annotators. We chose objects that were mentioned in all three sources (i.e., mentioned by at least one crowd worker and one study participant, as well as labeled by an annotator). We manually merged synonyms such as "jug" and "pitcher." For objects that had been identified by both annotators, we selected a bounding box for that object based on the intersection between the two bounding boxes. Figure 2 shows the bounding boxes generated by our two annotators. Overall, we identified 29 ground truth objects with an average of 3 objects per test image.

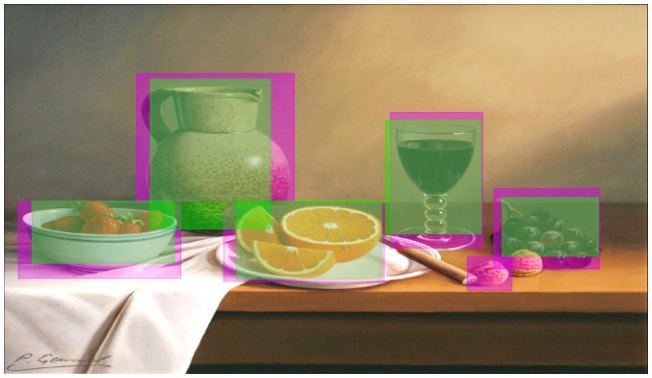

Figure 2. To collect ground truth data, two annotators drew bounding boxes for each object in an image. The canonical bounding box was the intersection of the two annotators' bounding boxes. Image Copyright Phillip Gerrard.

Findings

We analyzed the data collected during this activity to understand the relationship between eye gaze, speech, and objects in the image. In particular, we were interested to know when participants were gazing at our chosen ground truth objects, and how participants' behavior differed across the three conditions.

When Did Participants Look at Ground Truth Objects?

Activity 1. As there was no speech during Activity 1, we simply looked at what percentage of participants' gaze points landed on ground truth objects. Of the raw gaze points, 63% of these points were within the bounding boxes of the ground truth objects. Because this is raw gaze data, this data includes unconscious saccades. Looking at fixation data, 72% of

[3] https://annotorious.github.io

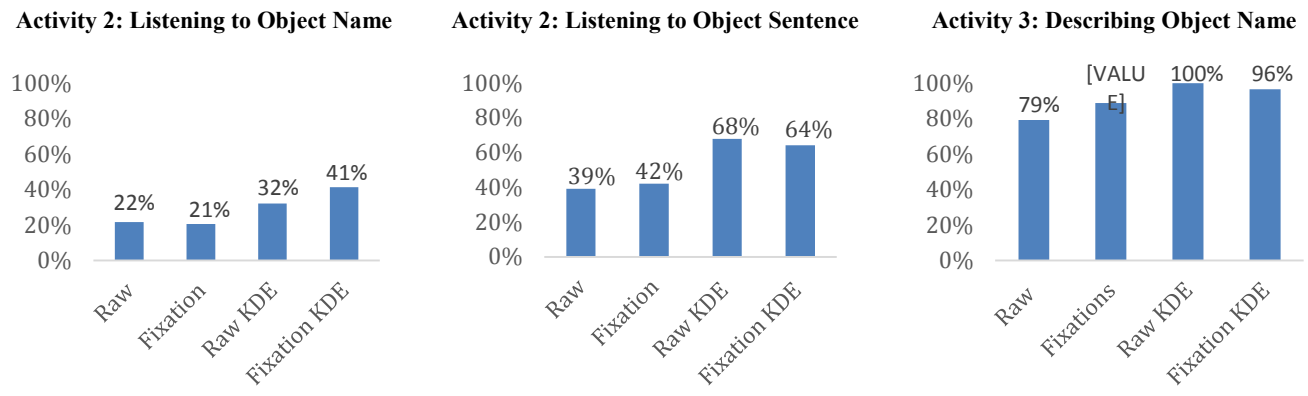

Figure 3. Raw gaze, fixation, raw KDE, and fixation KDE accuracy for the feasibility study. Higher Raw and Fixation % indicates that more of the gaze points were within the object during the sampled time window; higher KDE % indicate that the center of the gaze points was more often within the object during the time window. Overall, the user's gaze matched the named object most when the participant said the object's name, not when hearing a description of it.

fixations fell on a ground truth object. Participants averaged 137 fixations per image, and each fixation had an average duration of 283 milliseconds.

This data suggests that our labeled ground truth objects corresponded to the "interesting" parts of the image, and that participants tended to look at these objects without any explicit instruction to do so. In manually analyzing the gaze data, we found that participants sometimes fixated on other objects in the image, such as an artist's signature, that were not included in the set of ground truth objects. Thus, even when participants were not looking at the ground truth objects, they were sometimes gazing at more minor objects.

Activity 2. Participants gazed at the image while listening to a description of the image. Thus, we were able to examine whether participants looked at an object when hearing the name of the object (e.g., "oranges") and when hearing a sentence describing the object (e.g., "there are sliced oranges on a plate"). We were curious to know whether participants gazed at the object when hearing its name, hearing the sentence describing it, both, or neither.

Averaging across all (*participant, object*) pairs, 22% of raw gaze points and 21% of fixation gaze points were within an object's bounding box while hearing the object's name. When looking at the center of gaze, we found that participants centered their gaze on an object when hearing its name 32% of the time when calculating the center from raw data, or 41% when calculating the center from fixation data (Figure 3, left). Thus, participants directed much of their visual attention at an object when hearing its name, and often centered their gaze on that object when hearing its name.

We also examined how often participants' gaze matched an object while hearing the entire sentence describing that object. We found that examining gaze points over the entire sentence resulted in a greater overlap between language and gaze. Averaging across all (*participant, object*) pairs, 39% of raw gaze points and 42% of fixation gaze points were within an object's bounding box while hearing the object's sentence. When looking at the center of gaze, we found that participants centered their gaze on an object when hearing its sentence 68% of the time when calculating the center from raw data, or 64% when calculating the center from fixation data (Figure 3, center).

Activity 3. Finally, we examined agreement between gaze and language when the participant was describing the image themselves. We found a high level of agreement here: when describing an object and saying its name, 79% of raw gaze points (89% of fixation points) were within the object's bounding box. Examining the gaze center, we found that the participant's gaze was almost always centered on the object when describing it, 100% of the time when using raw KDE and 96% when using fixation KDE. This finding replicates prior work that noted that people gaze at an object while they are describing it [39].

Viability of this Approach (RQ1)

As in prior work [39], we found that participants tended to gaze at objects when hearing about that object or when describing it themselves. This was especially true when the participant was describing the image. Thus, we should be able to collect gaze and speech data as a sighted person describes an image and accurately label the objects they are describing. We use this approach in the development of our prototype system, EyeDescribe.

EYEDESCRIBE A

EyeDescribe is a system that takes data of the form we collected in the previous study (i.e., spoken descriptions and gaze data) and converts this data to an interactive representation. The EyeDescribe prototype runs on a Windows 10 laptop with a touch screen. Once an image is loaded in EyeDescribe, the user can explore the image using touch. As the user touches near the location of an object, EyeDescribe reads out a description of that object.

Dataset

For our initial prototype, we used the images and descriptions collected during Activity 2 from the previous study (eye gaze while listening to a description). This way we could assign sentence descriptions throughout the image rather than single words. For each image, we used the same description created by a Mechanical Turk crowd worker. We segmented each description into sentences and assigned each section to one of the ground truth objects used in the previous study. For each object, we chose its location by collecting the raw gaze data points collected while hearing that sentence, and then computing the raw KDE for that data. To ensure that several objects would fit on the tablet screen, each object was given a bounding box of the same size of 138 pixels by 138 pixels (1x1 inch).

User Interface

EyeDescribe uses a non-visual touch interface similar to Access Overlays [20]. The screen is divided into a series of Voronoi regions surrounding each labeled object [13,20]. The user can drag their finger around the screen; as they enter an object's Voronoi region, the system plays a unique note from the C-Major scale. The user can also perform a "find" gesture by double tapping the screen. When the user performs this gesture, the system describes the direction in which the user should move their finger, for example saying, "Move your finger 1 inch to the left and down 2 inches." When the user touches the object's center, the system reads the sentence describing that object.

To determine whether this spatial interface is useful, we compared this *spatial user interface* to a corresponding *linear user interface.* Both interfaces featured the same text content; however, in the linear UI, the image description is stored as a set of sentences. The user navigates through the sentences by swiping left and right, similar to the Apple VoiceOver screen reader on iOS devices.

USER STUDY 1

We conducted a study to compare linear and spatial descriptions, and the interfaces used to explore these descriptions.

Participants

We recruited 6 blind adult participants (2 female). Ages ranged from 23 to 64 with a mean age of 38. Four of the participants were congenitally blind, one has been blind for over nine years, and one participant had only light perception in one eye. All participants were familiar with accessible technologies and utilized screen readers daily.

Apparatus and Data

The EyeDescribe α prototype used a Microsoft Surface Pro 4 tablet, with a 12.3-inch, 3:2 ratio touch screen, running Windows 10. To answer RQ2 (*How do spatial captions generated by EyeDescribe compare to traditional text descriptions?*), we compared the spatial descriptions and linear descriptions as described above.

Procedure

The study lasted 30 minutes. This study used the same set of images as in the previous study. Half of the images were presented using the spatial interface and half were presented using the linear interface. The order of images was randomized. Interface conditions alternated every other image and were counterbalanced based on even or odd participant code.

For each image, participants were given as much time as they needed to explore the layout. Once they were done exploring the image, they moved on to the next image. After each interface condition, participants answered 5 Likert scale response questions (1=strongly disagree, 5=strongly agree):

1. **Object Presence:** I understand which objects are depicted in the image.
2. **Object Locations:** I understand where objects in the image are located.
3. **Ease of Use:** Exploring the image was easy.
4. **Enjoyability:** Exploring the image was enjoyable.
5. **Frustration:** Exploring the image was frustrating.

Finally, participants provided general feedback about their experience testing the prototype.

Findings

Feedback about the system focused on two main areas: 1) benefits of the spatialized captions; and 2) evaluation of the user interface. Overall, participants tended to agree that the spatialized captions contained useful information, but some had difficulty using the touch screen interface, which itself was an early prototype.

Table 1 shows the responses to the Likert-scale questions. Participants found the spatial user interface to provide a marginally better sense of the images' spatial layout, but also found it to be more difficult to use, less enjoyable, and more frustrating. Due to the small sample size, we did not conduct a statistical analysis of these data.

Category	Spatial UI	Linear UI
Object Presence	4.58	4.33
Object Locations	3.63	3.19
Ease of Use	2.75	4.31
Enjoyability	3.06	3.63
Frustration	3	1.75

Table 1. Responses to subjective questions (on a 5-point Likert scale, 1=strongly disagree, 5=strongly agree).

One major limitation of the spatial interface was the need to search the entire screen to locate the objects; participants needed to directly touch the object to hear its description. During the open-ended feedback session, P4 compared the spatial interface to playing hide and seek, and P6 said, "Spatial mode is very frustrating, it is very frustrating to wait for directions." The linear user interface was also more familiar to our participants as it was similar to the VoiceOver screen reader.

We analyzed participants' feedback about the spatial user interface and found a common set of concerns:

i1) It was difficult to find object descriptions;
i2) The sonified regions do not make sense due to their irregular shape boundaries;
i3) It was difficult to find the edges of the image;
i4) There was no way to toggle between interaction modes;
i5) The system did not provide enough feedback.

Despite the usability issues uncovered during the study, some participants noted benefits of the spatial user interface. P5 stated, "Overall I preferred the searching [spatial UI], I know the numbers I gave don't reflect that but I still prefer having that interaction instead of just being told. [Spatial UI] would be better for getting to engage with artwork instead of being lectured at." P2 said, "I liked the ones[images] in the spatial mode the most. You could tell where things were in relation to one another … If someone has more vision than I do, then it would allow me to communicate better with sighted people."

EYEDESCRIBE β

Following our first user study, we updated the EyeDescribe prototype to address participants' usability issues and to better support spatial exploration of the image.

Updated Image Set

During the first study, some of the images in our image set contained multiple instances of similar objects. Specifically, some images featured multiple copies of the same object (giraffes, squares, and dogs, respectively). Participants were able to locate one instance of these objects, but were mostly unaware of the additional instances of each object, as EyeDescribe attempts to find a single location for each type of object. Also, some images in the initial data set had many objects in them, while others had very few. To remove ambiguous duplicate objects, and to equalize the number of findable objects in each image, we selected a new image set, and collected descriptions and eye gaze data for each image. This set contained 4 images from 3 categories: landscapes, portraits, and still-life paintings.

Once again, we recruited gazers from the local campus with normal or corrected-to-normal vision. We recruited 5 participants (3 female) with a mean age of 25 (range 23–30). As we found the image description task to be especially accurate, participants in this group only described images in their own words and did not complete the other activities.

User Interface Improvements

To address the usability issues uncovered in the first study, we made several improvements to the touch screen interface. First, we added a laser-cut tactile border to the tablet screen to make it easier to find the screen and navigate along its edges (Figure 6). We also provided an enhanced set of gestures partly inspired by the Access Overlays [20], with an emphasis on providing users more information about the image and its spatial layout.

Single Tap: Read Image Caption

The user can perform a single tap to start playing a description and another single tap to pause the description. The description can be played as many times as the user prefers. This was most similar to the linear interface in the previous prototype.

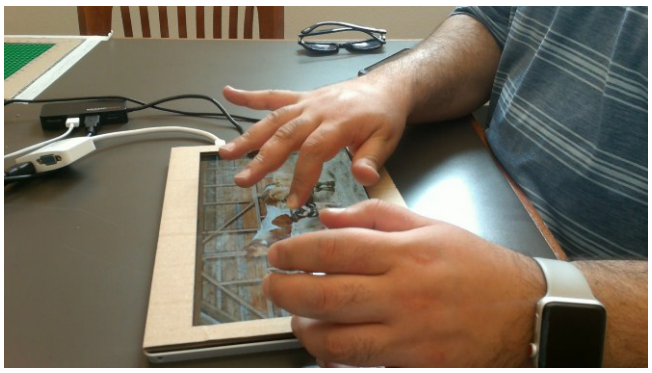

Figure 4. Participant interacts with the EyeDescribe prototype.

Double Tap: List Common Objects

The user can double tap the screen to hear a list of 10 nouns that were most commonly mentioned in our gazers' image descriptions. For example, for a still life image, EyeDescribe may read, "Objects associated with image: glass, wine, bowl, grapes, table, liquid, apple, cloth, apples, pitcher." We found that this set of words described the most important objects in our test image set.

Two-Finger Tap: Search

The user can tap the screen with two fingers to perform an interactive search, as in Access Overlays [20]. The user taps the screen and then says the name of the object; the system then provides interactive directions to guide the user's finger to the named object. While in the previous prototype the user needed to repeatedly tap the screen, here the user can drag their finger around until they locate the object.

Touch and Drag: Scan

The user can drag a single finger over the screen. As the user moves from the region around one object to the region around another object, the system will say the name of the new object. While in the previous prototype the user needed to accurately pinpoint the object to hear its name, here the user can drag their finger around to get an overview of the image layout.

USER STUDY 2

We conducted another user study to evaluate whether the improved EyeDescribe β would enable participants to more efficiently explore images.

Participants

We recruited 7 blind participants (2 female). Four of the seven participated in *User Study 1*. Ages ranged from 29 to 66 with a mean age of 40. P1 has been blind for over 10 years. P2 has been blind for over 5 years. The other participants were congenitally blind. The participants were

familiar with accessible technologies and used screen readers daily.

Procedure

The study lasted 90 minutes. During the study, participants interacted with 10 images from the test set. There were three parts to the study:

Activity 1: Training

Participants explored a sample image using the touch screen interface. Participants were able to use all four of the interface modes.

Activity 2: Explore

The participant explored images using the EyeDescribe interface. For this activity, participants explored three images, featuring one example from each category (landscape, still-life, and portrait). The order of the images was the same in this task across participants. Participants were asked to think aloud while exploring the images and to take as much time as they wished to explore the images. Participants were able to use all four of the interface modes.

Activity 3: Recall

Participants explored eight additional images. These images were the same for all participants but were presented in random order. For four images, participants were able to use all four of the interface modes; for the other four images, they were only able to use the non-spatial interface modes (*Read Image Caption* and *List Common Objects*). The order of the modes was counterbalanced.

Once the participant felt they had adequately explored the image, they were then asked to recreate the layout of the image. The participant was presented with a Lego grid that was approximately the size of the tablet screen. For each image, the researcher named one object (same object for each participant) and asked the participant to place a Lego brick on the grid corresponding to that object's location. The goal of this activity was to determine whether the spatial user interface helped participants to understand the image layout.

Findings

We analyzed the completed Lego grids and measured the distance (in inches) between the object's on-screen location and the position that the participant chose. Table 2 shows the average distance for images using the spatial interface (all four interface modes) and the non-spatial interface (*Read Image Caption* and *List Common Objects* modes).

Although the average error distance was less when using the spatial interface, a one-way repeated measures ANOVA [$F(1,27)=3.4694$, $p=0.0734$] did not show this difference to be significant. This may be an artifact of the small sample size. Cohen's d (0.40) was calculated for effect size.

	Sum	Mean	VAR	p-value
Spatial	53.18	1.90	2.26	0.0734
Non-Spatial	70.99	2.26	2.78	

Table 2. Distance (in inches) between on-screen location and the participant's estimated location. Participants were more accurate when using the spatial interface, although the difference was not statistically significant.

Qualitative feedback about EyeDescribe β was more positive than the feedback about EyeDescribe α. Following the improvements to the user interface, participants seemed more able to search and understand images. For example, P3 said, "The worst was when there was no explore mode [non-spatial interface] and just going off of people's description."

Participants were asked what they liked most about exploring the images. P1 said, "It allowed me to get a mental image drawn in my brain … I like the ability to explore and look for things." When asked to compare the various modes of the prototype, P2 said "They are all important in terms of providing the information we need." Despite the potential advantages of a spatial user interface, two participants noted that their favorite feature was the plain text description, although two other participants named that their least favorite feature.

When asked how this system might be useful to them, P2 said, "Obviously describing art." Five out of seven participants mentioned that the system would be useful for maps; four out of seven participants said they might use EyeDescribe to look at pictures. Participants also mentioned museum exhibits, games, and documents as potential applications of this system.

DISCUSSION & FUTURE WORK

In this work, we have explored the creation of spatially-captioned interactive graphics using use-driven accessibility. We believe that this work demonstrates that through capturing the behavior of consumers as they interact with media, we can bootstrap the creation of accessible media.

EyeDescribe could be extended to adapt its descriptions based on behavioral data. For example, EyeDescribe might present more detailed descriptions of objects that sighted users spend a lot of time looking at. Alternately, the user could query an image based on behavioral data, such as searching for the most popular objects, or could filter the description data so that they can switch between object labels, descriptions of color and texture, or subjective comments about an image. Future versions of EyeDescribe could allow a user to follow along with sighted peers in real-time, or to switch between different gazers to understand how individuals experience a work of art differently. While EyeDescribe currently expects users to talk about an image

as they look at it, future versions could further reduce the labeler's workload by recording and processing the natural discussions that occur around a piece of art.

Use-driven accessibility might be generalizable to other use contexts and disabilities. For example, tracking movements through a building could be used to identify accessible pedestrian paths, and sets of items purchased in a store could be used to provide suggestions for people with cognitive disabilities while shopping. This approach could even be considered a form of social accessibility [25], whereby members of the community can contribute to the accessibility of some shared resource, albeit one in which no explicit extra work is required from the community members.

CONCLUSION

In this paper, we introduced the concept of use-driven accessibility and show how it can be used to label artistic images. This approach attempts to capture what a sighted viewer intuitively finds interesting in an image and makes this information accessible with little or no work from the sighted labeler. Our studies showed that the combination of gaze and speech is sufficient for labeling images and that automatically generated spatial captions can lead to improved understanding over plain text captions. We believe that understanding how people engage with a piece of art can provide unique insights into how to make that art more accessible to everyone.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grant IIS-1652907. Any opinions, findings, conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect those of the National Science Foundation.

REFERENCES

[1] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. *Proceedings of the 2004 conference on Human factors in computing systems (CHI '04)*. 319–326. DOI:https://doi.org/10.1145/985692.985733

[2] Fernando Alonso, José Luis Fuertes, Ángel Lucas González, and Loïc Martínez. 2010. On the testability of WCAG 2.0 for beginners. *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A '10)*1. DOI:https://doi.org/10.1145/1805986.1806000

[3] Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight: Making Web Images Accessible. *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility (Assets '06)*. 181. DOI:https://doi.org/10.1145/1168987.1169018

[4] Stefania Bocconi, Silvia Dini, Lucia Ferlino, Cristina Martinoli, Michela Ott, H Jürgensen, C Power, Rabia Jafri, Syed Abid Ali, Saudi Arabia, Giovanni Fusco, San Francisco, Valerie S Morash, Kevin Allain, Mick Van Gelderen, Olivier Hokke, Miguel Oliveira, Richard C Hendriks, Ben Kybartas, Lisa Tebo, and Fouzia Khursheed Ahmad. 2014. The Tactile Graphics Helper : Providing Audio Clarification for Tactile Graphics Using Machine Vision Categories and Subject Descriptors. *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*: 491–500. DOI:https://doi.org/10.1145/2700648.2809868

[5] Craig Brown and Amy Hurst. 2012. VizTouch. *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction (TEI '12)*. 131. DOI:https://doi.org/10.1145/2148131.2148160

[6] G T Buswell. 1935. *How people look at pictures: a study of the psychology and perception in art.* Univ. Chicago Press, Oxford, England.

[7] Stephen H. Choi and Bruce N. Walker. 2010. Digitizer auditory graph: making graphs accessible to the visually impaired. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 3445-3450. DOI:https://doi.org/10.1145/1753846.1753999

[8] Andrew T. Duchowski. 2017. *Eye Tracking Methodology: Theory and Practice*. (3rd. ed.). Springer, GA

[9] Polly K. Edman. 1992. *Tactile graphics*. American Foundation for the Blind, New York, NY

[10] Jim Edwards. 2014. PLANET SELFIE: We're now Posting a Staggering 1.8 Billion Photos Every Day. *Business Insider Autralia*.

[11] Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, Andrew Zisserman. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *Journal of Computer Vision* 88. 303–338. DOI:https://doi.org/10.1007/s11263-009-0275-4 1.

[12] Chuu-hai Fan. 2004. Driver Fatigue Detection Based. *Proceedings of the 2004 IEEE, International Conference on Networking, Sensiing & Control*. 7–12. DOI:https://doi.org/10.1109/ICNSC.2004.1297400

[13] Steven Fortune. A sweepline algorithm for Voronoi diagrams. Algorithmica 2, 1 (1987), 153-174.

[14] Albert F. Fuchs. 1971. *The Saccadic System*. The Control of Eye Movements.

[15] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Paper 518, 11 pages. DOI: https://doi.org/10.1145/3173574.3174092

[16] Edmund Huey. 1908. *The Psychology and Pedagogy of Reading*. The Macmillan Company.

[17] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J Donnelly, and Sidney K D 'Mello. 2016. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. *Proceedings of the 9th International Conference on Educational Data Mining. International Educational Data Mining Society (EDM '16)*. 86–93.

[18] Chandrika Jayant, Matt Renzelmann, Dana Wen, Satria Krisnandi, Richard Ladner, and Dan Comden. 2007. Automated Tactile Graphics Translation : In the Field. 75–82.

[19] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 333-342. DOI:https://doi.org/10.1145/1866029.1866080

[20] Shaun K. Kane, Meredith Ringel Morris, Annuska Z. Perkins, Daniel Wigdor, Richard E. Ladner, Jacob O. Wobbrock, and Meredith Ringel Morris. 2011. Access Overlays: Improving Non-Visual Access to Large Touch Screens for Blind Users. *Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11)*. 273–282. DOI:https://doi.org/10.1145/2047196.2047232

[21] Shaun Kane, Brian Frey, and Jo Wobbrock. 2013. Access Lens: A Gesture-Based Screen Reader for Real-World Documents. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 347–350. DOI:https://doi.org/10.1145/2470654.2470704

[22] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze Embeddings for Zero-Shot Image Classification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 4525-4534

[23] Oleg V. Komogortsev, Sampath Jayarathna, Cecilia R. Aragon, and Mechehoul Mahmoud. 2010. Biometric identification via an oculomotor plant mathematical model. *In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. ACM, New York, NY, USA, 57-60. DOI:http://dx.doi.org/10.1145/1743666.1743679

[24] Maurice Kunder. 2016. The Size of the World Wide Web. *WorldWideWebStats*. Retrieved January 8, 2018 from https://worldwidewebsize.com

[25] Sam Lucente, Steve Sato, Deborah Mrazek, and Douglas Meyer. 2010. Design Thinking to Make Organization Change and Development More Responsive. *DMI Review* 21, 2. (June 2010) 44-52. DOI:https://doi.org/10.1111/j.1948-7169.2010.00064.x

[26] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 5988–5999. DOI:https://doi.org/10.1145/3025453.3025814

[27] William D. Marslen-Wilson. 1984. Function and process in spoken word recognition: A tutorial review. *Attention and performance: Control of language processes*. 125–150.

[28] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Paper 59, 11 pages. DOI: https://doi.org/10.1145/3173574.3173633

[29] Petrie, Helen, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII '05)* 71. July 2005.

[30] Kyle Rector, Neel Joshi, and Meredith Ringel Morris. Eyes-Free Art : Proxemic Audio Interfaces for Blind and Low Vision Art Exploration. *Interactive, Mobile, Wearable and Ubiquitous Technologies*: 1–21.

[31] Andreas Reichinger, Anton Fuhrmann, Stefan Maierhofer, and Werner Purgathofer. 2016. Gesture-Based Interactive Audio Guide on Tactile Reliefs. *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. 91–100. DOI:https://doi.org/10.1145/2982142.2982176

[32] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the symposium on Eye tracking research & applications (ETRA '00)*. 71–78. DOI:https://doi.org/10.1145/355017.355028

[33] Andrew Salway and Burton Bradstock. 2007. A Corpus-Based Analysis of Audio Description. In *Media for All*. Brill, Rodopi, 2007.

[34] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 771-780. DOI:http://dx.doi.org/10.1145/1124772.1124886

[35] Lei Shi, Yuhang Zhao, and Shiri Azenkot. 2017. Markit and Talkit: A Low-Barrier Toolkit to Augment 3D Printed Models with Audio Annotations. *In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York,

NY, USA, 493-506. DOI:https://doi.org/10.1145/3126594.3126650

[36] Joel Snyder. 2005. Audio description: The visual made verbal. *International Congress Series* 1282: 935–939. DOI:https://doi.org/10.1016/j.ics.2005.05.215

[37] Abigale Stangl, Chia-Lo Hsu, and Tom Yeh. 2015. Transcribing Across the Senses: Community Efforts to Create 3D Printable Accessible Tactile Pictures for Young Children with Visual Impairments. *In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 127-137. DOI:https://doi.org/10.1145/2700648.2809854

[38] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. 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* (CSCW '17), 1180–1192. https://doi.org/10.1145/2998181.2998364

[39] K Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, Tamara L. Berg. 2013. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)* 739-746.

[40] 2008. Web content accessibility guidelines (WCAG) 2.0. *World Wide Web Consortium*.